# METHODS AND APPLICATIONS FOR SPACE-TIME DATA

BY

LYNDSAY SHAND

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Associate Professor Bo Li, Chair
Professor Jeff Douglas
Assistant Professor Stéphane Guerrier
Clinical Associate Professor Marilyn O'Hara Ruiz

# Abstract

Spatial and spatio-temporal data are presented in a variety of forms and require a unique set of techniques to analyze. The goal of such analyses is often to estimate the spatial and/or temporal dependency structures of the underlying random field. This estimation in turn can then be used to make inference about the underlying random process. Recurring challenges with spatial data include a lack of multiple realizations of the process, e.g. a lack of replicates and the estimation of dependency structures given this difficulty. In this work, I contributed to solving this problem based on both geostatistical and areal data using likelihood and Bayesian methods respectively.

A nonstationary spatio-temporal model is proposed which applies the concept of the dimension expansion method in Bornn et al. (2012). The estimation of this model is investigated and simulations are conducted for both separable and nonseparable space-time covariance models. The model is also illustrated with wind speed and streamflow datasets. Both simulation and data analyses show that modeling nonstationarity in both space and time can improve the predictive performance over stationary covariance models or models that are nonstationary in space but stationary in time.

In demand of predicting new HIV diagnosis rates based on publicly available HIV data that is abundant in space but has few points in time, a class of spatially

varying autoregressive (SVAR) models compounded with conditional autoregressive (CAR) spatial correlation structures is proposed. The copula approach coupled with a flexible CAR formulation are employed to model the dependency between adjacent counties. These models allow for spatial and temporal correlation as well as space-time interactions and are naturally suitable for predicting spatio-temporal disease data that feature such a data structure. The models also allow us to estimate the spatially varying evolution pattern of the disease. We apply the proposed models to HIV data over Florida, California and New England states and compare them to a range of linear mixed models that have been recently popular for modeling spatio-temporal disease data. The results show that for such data our proposed models outperform the others in terms of prediction.

# Acknowledgements

I want to thank my advisor Dr. Bo Li for her guidance and patience and my family for all of their support. I also want to thank my committee members for their time and Drs. Trevor Park and Dolores Albarracín for their collaboration and helpful advice.

# Table of Contents

# Chapter 1

# Introduction

Spatial data is most commonly used in fields such as climatology, epidemiology, geology, agriculture, and environmental sciences. Any dataset which can be mapped, that is, has geographical coordinates associated with each observed measurement, is considered spatial data. Formally, we denote a spatial process, or a spatial random field, on a $d$-dimensional domain $D$ as $\{Y(s), s \in D \subset \mathbb{R}^d\}$ where $Y$ is the observed variable or measurement and $s$ is a vector of coordinates of dimension $1 \times d$. This collection of measurements is considered a single sample from the underlying spatial process. For example, air temperature measured at weather stations across the United States at a single time point or the number of human disease cases in each county across Illinois for a given year. Temperature across the U.S is known to have spatial correlation as it tends to increase the closer you are to the equator. The number of disease cases, especially of those which are transmitted through contact, are usually more similar among neighboring counties than between those that are geographically far apart.

As there are a wide variety of fields in which spatial analysis is applicable, there are

a plethora of spatial data types and structures. Literature typically classifies spatial data in three main categories: geostatistical data, lattice or areal data, and point data. Geostatistical data refers to the spatial process of a random variable observed over a continuous space such as air temperature across the US. Each data point is associated with a specific geolocation such as longitude and latitude coordinates. The focus of analysis with geostatistical data is in modeling the spatial covariance structure and in predictions at unobserved locations. Lattice data refers to data which is typically aggregated at fixed locations, which are a countable collection of bounded spatial regions such as state counties or census tracts. A single summary statistic is associated with an entire region rather than with a specific set of coordinates, for example the total number of disease cases by county. Since there are no precise geolocations associated with regions, the focus of analysis is typically in modeling the spatial trend rather than prediction. Given the nature of lattice data, it can also be helpful to identify clusters, i.e. neighboring regions with similar values, and to associate such clusters with other datasets, say disease cases with environmental risk factors. Finally, point data refers to a collection of locations of an event of interest. The response variable at each location is binary, e.g. whether the event occurred or not. In contrast to geostatistical data, the attribute observed at each location is unimportant and rather it is the locations of the events that are treated as the outcome of a random process. The goal of analysis is to determine whether the dispersion of locations are clustered or purely random. Note that point data can be treated as lattice data by grouping locations into regions and recording the number of event occurrences within each region. For more detailed information on point pattern analysis, see *Mapped Point Patterns* in Schabenberger and Gotway (2004). From here forward, this dissertation will focus on the analysis of geostatistical data and lattice data.

As with traditional statistical methods, inference in spatial analysis aims to estimate the underlying process model and predict unobserved data. Modeling a spatial process involves estimating the spatial correlation structure of the underlying process so inference can then be made at unobserved locations, which is important when we cannot obtain an exhaustive spatial sample. For example, it is unfeasible to place a weather station at every point across the United States. By understanding the spatial structure of temperature ranges across the United States using the existing weather stations, we can estimate a prediction surface and predict temperatures at any of the station-less locations.

Important features of $Y(s)$, and of main interest to estimate, are its mean and covariance functions. The former gives information about the large scale trend of $Y$ while the latter, small scale changes. Suppose we observe a sample of the process $Y$ of size $n$ such that we have observations $Y(s_1), Y(s_2), ..., Y(s_n)$, with $E[Y(s_i)] = \mu(s_i)$ and $Var[Y(s_i)] = \sigma^2$. The mean function $\mu$ can either be a constant or expressed as a function of covariates as seen in (1.0.1). We define the spatial covariance between two locations $s_i$ and $s_j$ as

$$cov[Y(s_i), Y(s_j)] = E[(Y(s_i) - \mu)(Y(s_j) - \mu)]$$

and spatial correlation as

$$corr[Y(s_i), Y(s_j)] = \frac{cov[Y(s_i), Y(s_j)]}{\sigma^2}.$$

A basic model for $Y(s)$ can be written as

$$Y(s) = X\beta + \nu(s) \tag{1.0.1}$$


Figure 1.1: Locations of Illinois weather stations. *http://www.isws.illinois.edu*

with spatial random effect $\nu(s)$ such that $Cov[\nu(s)] = \Sigma$ where $\Sigma_{i,j} = Cov[Y(s_i), Y(s_j)]$. Here, the mean function $\mu(\boldsymbol{s})$ is expressed as $X\beta$ where $X$ is an $n \times p$ matrix of known $p$ covariates. Ignoring spatial correlation when modeling data that has known spatial dependency can lead to model mispecification. For example, if there is spatial correlation but we assume a typical linear regression model with errors $\boldsymbol{\nu} \overset{iid}{\sim} (0, \sigma^2)$, it will lead to incorrect variance estimation for $\hat{\beta}$ and prediction. Thus, when we have spatially correlated errors, it is important to estimate $\Sigma$, which describes the spatial dependencies of the data.

## 1.1 Geostatistics

A spatial process $Y(s)$, observed at locations $s \in \mathbb{R}^d$, is considered geostatistical data if $s$ is allowed to vary continuously over domain space $D$, a fixed subset of $\mathbb{R}^d$. Theoretically, there are an infinite number of possible observations of $Y$ between

two locations $s_i$ and $s_j$. Practically, since we cannot sample a continuous space exhaustively, we only observe a finite sample. However, we can build a model for the continuous spatial surface across $D$, i.e. a mapping of $Y$, to describe the underlying structure of the spatial process. The primary goal of modeling a spatial random field (SRF) $Y(s)$ is twofold: (1) understand the relationship between any given pair of locations $s_i$ and $s_j$; (2) predict the spatial process at an unobserved location $s_0$ or over a region of interest B, $\int_B Y(s)ds$. Weather data is often used as examples for geostatistical data. For example, if we are interested in wind speed throughout Illinois, we can use the observed the wind speed at all weather stations in Illinois (Fig. 1.1) to estimate the spatial covariance structure and then make prediction at unobserved locations (where there are no weather stations) across the continuous space of Illinois. Figure 1.2 shows a predicted surface for maximum wind speed in miles per hour over Illinois on April 17,2010.

Before looking at examples of spatial models and methods of estimation and prediction, it is important to consider the properties of the SRF $Y(s)$ and common assumptions made in modeling it's covariance structure. The most common modeling assumptions made about a SRF are second-order stationarity (SoS) and intrinsic stationarity (IS). Note that in theory stationarity is not always an appropriate assumption, especially in environmental processes. In practice, it is common to assume SoS because it allows us to have replicates to estimate the covariance function. A random field is said to be SoS if it has a constant mean and constant variance across the domain and it's covariance structure only depends on the distance $h$ between locations, i.e. $E[Y(s+h) - Y(s)] = 0$, $Var[Y(s)] = Var[Y(s+h)]$, and we can write $Cov[Y(s+h), Y(s)] = C(h)$. The latter states that the covariance between locations does not depend on the exact coordinates $s_i$ and $s_j$, but only on the distance between them, $h = s_i - s_j$. If further, $C(h)$ is a function only of $||h||$, then the SRF is called

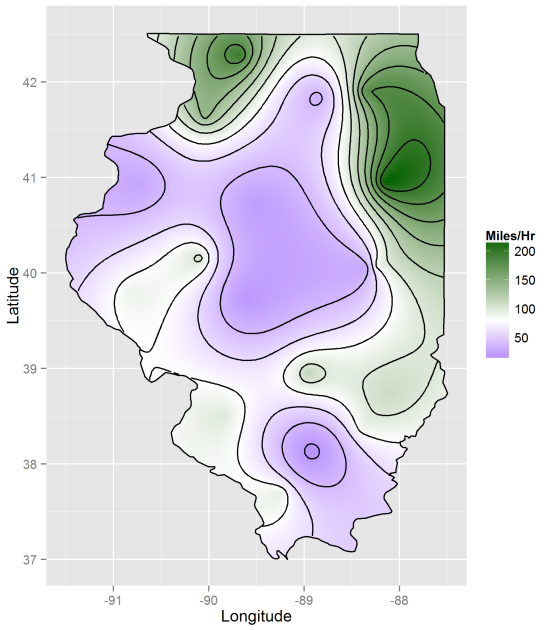


Figure 1.2: Kriging predictions for maximum wind speed in miles per hour across Illinois.

isotropic, or invariant to the direction of the field. Anisotropy occurs when $C(h)$ is a function of both the magnitude and direction of $h$. If the field is IS, then we only require $Var[Y(s+h) - Y(s)] = 2\gamma(h)$ to be satisfied where $2\gamma(h)$ is called the variogram function of $Y(s)$ and $\gamma(h)$, the semi-variogram function. Notice that if the process was not IS, $\gamma$ would not exist.

If a process is second-order stationary, one could choose to work with either the covariance function $C(h)$ or the semi-variogram $\gamma(h)$. To be a valid covariance function, $C(h)$ must be nonnegative-definite, namely for any real numbers $\{a_i,\ i = 1, 2, ..., n\}$ and finite locations $\{s_i, i = 1, 2, ..., n\}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(h_{i,j}) \geq 0,$$

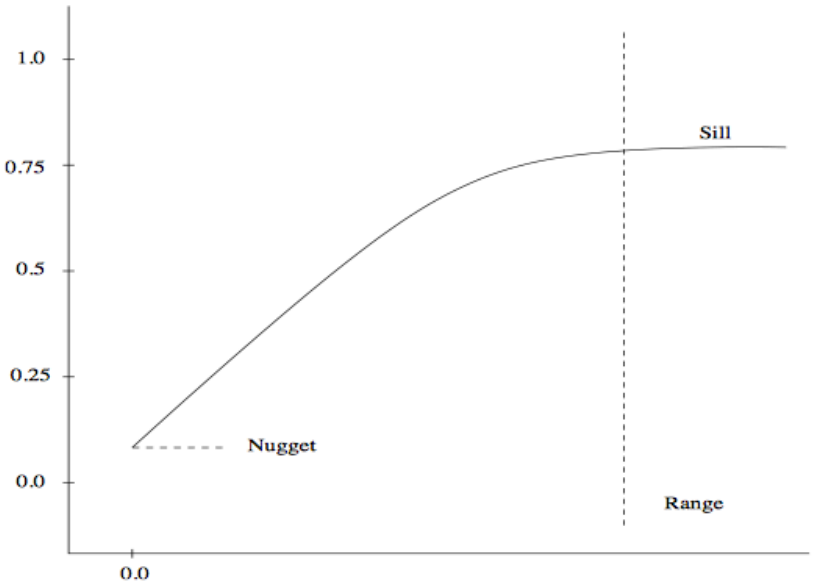


Figure 1.3: Range, sill and nugget parameters of a variogram model.

where $h_{i,j} = s_i - s_j$. A valid variogram function must be conditionally negative-definite, namely for any real numbers $\{a_i i, 1, 2, ..., n\}$ such that $\sum_{i=1}^{n} a_i = 0$ and finite locations $\{s_i, i = 1, 2, ..., n\}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(h_{i,j}) \leq 0.$$

If $Y(s)$ is SoS, then these conditions are the same as a valid covariance function implies a valid corresponding variogram function.

Let $t = ||h||$ and consider the isotropic exponential covariance model:

$$C(t) = \tau^2 \mathbb{1}_{t=0} + \sigma^2 e^{-\phi t},$$

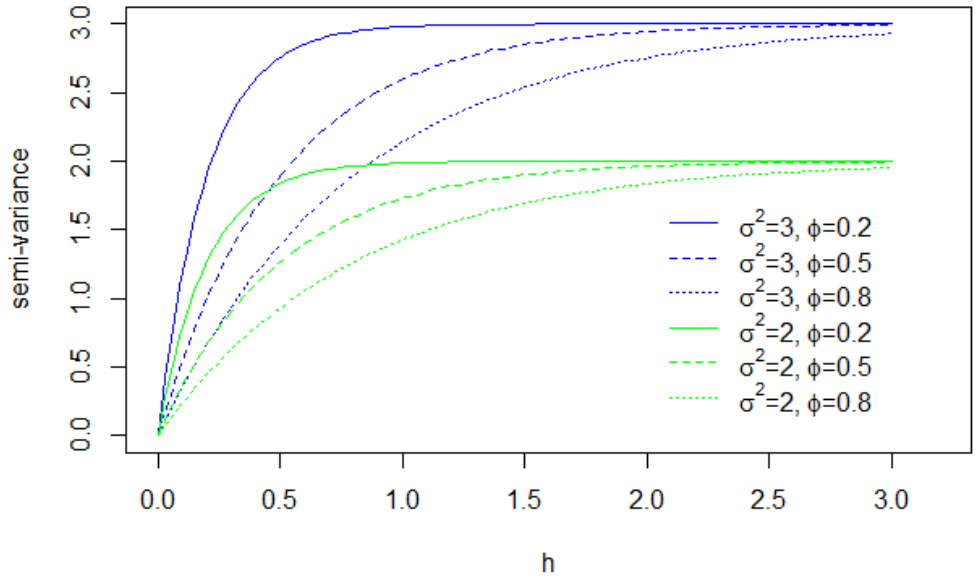


Figure 1.4: Exponential Variogram Model

with corresponding variogram model

$$\gamma(t) = C(0) - C(t) = \tau^2 \mathbb{1}_{t=0} + \sigma^2\big(1 - e^{-\phi t}\big).$$

where $\tau^2 = lim_{t\to 0}\gamma(t)$ is the nugget effect, $\tau^2 + \sigma^2 = lim_{t\to\infty}\gamma(t)$ is called the sill and $\sigma^2 = Var[Y(s)]$ is the variance of the process. See Fig. 1.3 for an illustration of the nugget and sill. The range parameter $\phi$ measures how quickly the correlation decays as $t$ increases. A larger $\phi$ indicates a slower decay. The lag distance $t^*$ such that $C(t^*) = 0$ is termed the range of the process (Fig. 1.3) and pairs of locations further than $t^*$ distance apart are considered negligibly correlated. If the process decays slow enough such that $\gamma(t)$ reaches the sill only asymptotically, the distance $t$ where $\gamma(t)$ increases to 95% of $\tau^2 + \sigma^2$ is termed the practical range. Figure 1.4 shows how the

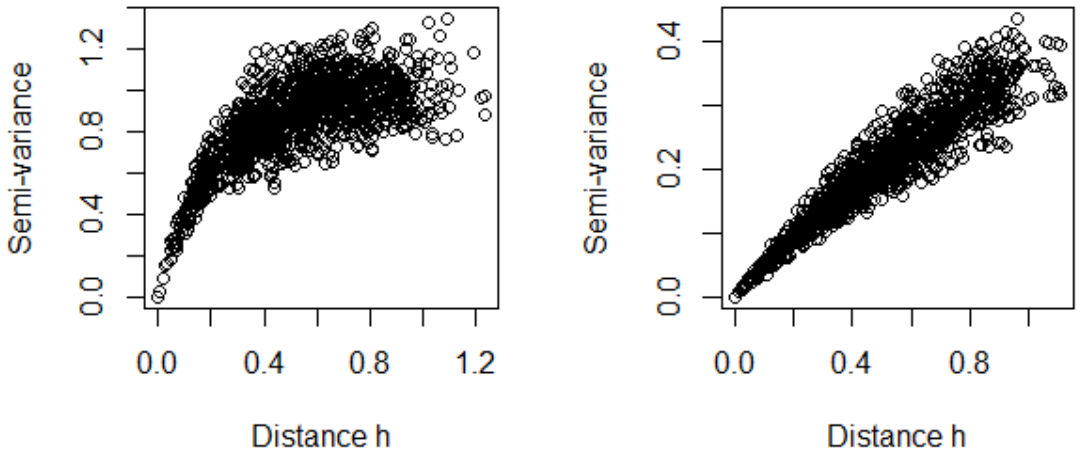


Figure 1.5: Empirical semi-variogram with $\sigma^2 = 1$ and exponential correlation structure with moderate, $\phi = 5$ (left) and strong, $\phi = 0.5$(right) correlation.

shape of an exponential variogram model with $\tau^2 = 0$, changes with different variance and range parameters, $\phi = 0.2, 0.5, 0.8$ indicating strong, moderate and weak spatial correlation respectively. When the spatial correlation is strong, the semi-variance decays at a faster rate and has a shorter range than when the correlation is weak. We can generally express the relationship between $C(h)$ and $\gamma(h)$ of an isotropic process as:

$$\gamma(t) = \begin{cases} \tau^2 & if\ t = 0 \\ \tau^2 + \sigma^2(1 - V(t)) & if\ t > 0 \end{cases}$$

where $V(t) = C(t)/\sigma^2$ is the called the spatial autocorrelation function or correlogram.

Estimation of variogram and covariance parameters can be done in many ways. Classically, for a general IS process, the following empirical estimator was proposed

by Matheron:

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} (Y(s_i) - Y(s_j))^2,$$

where $N(h)$ is the number of location pairs which are a separated by a distance of $h$. A plot of $\hat{\gamma}(h)$ against $||h||$ is called the empirical semi-variogram cloud. Figure 1.5 shows the empirical semi-variogram clouds of 100 simulations, $N(h) = 100$, from a Gaussian process at $n = 50$ locations on a spatial domain of $(0, 1) \times (0, 1)$ with moderate,$\phi = 5$, and strong,$\phi = 0.5$, exponential correlation structures. Cressie (1991) shows that this empirical estimator is an unbiased estimate of the theoretical variogram.

In practice, a more popular choice of estimators is the maximum likelihood estimator (MLE). Consider the common modeling framework in (1.0.1) and assume $Y(s)$ is Gaussian random process such that $\nu(s) \sim N(0, \Sigma(\theta))$. We can then derive the MLE estimates $\hat{\beta}$ and $\hat{\theta}$ by minimizing the negative log likelihood function:

$$l(\beta, \theta) = \frac{n}{2} log(2\pi) + \frac{1}{2} log(|\Sigma(\theta)|) + \frac{1}{2}(Y - X\beta)^T \Sigma(\theta)^{-1}(Y - X\beta)$$

. The minimization can be achieved using optimization algorithms such as Nelder-Mead or Broyden, Fletcher, Goldfarb, and Shanno (BFGS) method. We can reduce the number of free parameters in $l(\beta, \theta)$ by instead minimizing the profile log likelihood $l(\theta)$. By writing $\Sigma(\theta) = \sigma^2 V(\theta)$, the negative log-likelihood can be expressed as:

$$l(\beta, \sigma^2, \theta) = \frac{n}{2} log(2\pi) + \frac{n}{2} log(\sigma^2) + \frac{1}{2} log(|V(\theta)|) + \frac{1}{2\sigma^2}(Y - X\beta)^T V(\theta)^{-1}(Y - X\beta). \tag{1.1.1}$$

It can be shown that the MLE of $\beta$ is

$$\hat{\beta} = (X^T V(\theta)^{-1} X)^{-1} X^T V(\theta)^{-1} Y,$$

and given $\hat{\beta}$, the MLE of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{(Y - X\hat{\beta})^T V(\theta)^{-1} (Y - X\hat{\beta})}{n}. \tag{1.1.2}$$

Plugging these estimates into (1.1.1), we have the profile log likelihood:

$$l(\theta) = \frac{n}{2} log(2\pi) + \frac{n}{2} log\big(\frac{1}{n}(Y - X\hat{\beta})^T V(\theta)^{-1} (Y - X\hat{\beta})\big) + \frac{1}{2} log(|V(\theta)|) + \frac{n}{2}.$$

After removing constants, the profile log likelihood reduces to:

$$l(\theta) = \frac{n}{2} log\big(\frac{1}{n}(Y - X\hat{\beta})^T V(\theta)^{-1} (Y - X\hat{\beta})\big) + \frac{1}{2} log(|V(\theta)|). \tag{1.1.3}$$

Using the profile likelihood, we first estimate $\hat{\theta}$ and $\hat{\beta}$, then we obtain $\hat{\sigma}^2$ by (1.1.2). This reduces the number of unknown parameters involved in the likelihood function.

We have seen that with a finite observed sample $Y(s_1), Y(s_2), ..., Y(s_n)$ of a spatial process, we can estimate the true underlying model of the SRF over the continuous domain $D$. We can then construct a prediction surface $\hat{Y}(s)$ based on this model over $D$ and predict at an unobserved location $s_0 \in D$. Consider the model:

$$Y(s) = \mu(s) + \nu(s),$$

where $\nu(s) \sim (0, \Sigma))$ is an intrinsic stationary process and assume $\mu(s)$ is known. The "best" predictor $\hat{Y}(s_0)$ will ideally minimize the squared error loss or mean squared

prediction error (MSPE):

$$MSPE(\hat{Y}(s_0)) = E\big[(\hat{Y}(s_0) - Y(s_0))^2\big]$$

where both $\hat{Y}(s_0)$ and $Y(s_0)$ are random. Then it is clear that

$$\hat{Y}(s_0) = E[Y(s_0)|Y(s_1), Y(s_2), ..., Y(s_n)]$$

is the best predictor. It is traditionally assumed that $Y(s_0)$ is a linear function of $Y(\boldsymbol{s})$, i.e.

$$Y(s_0) = \lambda_0 + \sum_{i=1}^{n} \lambda_i Y(s_i)$$

where $\lambda_i, i = 0, 1, ..., n$ are unknown constants. To find the best linear unbiased predictor (BLUP) $\hat{Y}(s_0)$, it is sufficient to find $\boldsymbol{\lambda}$ that minimizes the MSPE. With some simple algebra, it can be shown that $\boldsymbol{\lambda} = \Sigma^{-1}\sigma$ where $\sigma = Cov[Y(s_0), Y(s_1), Y(s_2), ..., Y(s_n)]$. Thus,

$$Y(s_0) = \mu(s_0) + \sigma^T\Sigma^{-1}(Y(\boldsymbol{s}) - \mu(\boldsymbol{s}))$$

and $\hat{Y}(s_0) = \hat{\boldsymbol{\lambda}}Y(\boldsymbol{s})$ is the BLUP and is called the simple kriging predictor with prediction variance

$$\sigma^2(s_0) = C(0) - \sigma^T\Sigma^{-1}\sigma. \tag{1.1.4}$$

If $\mu(s)$ is unknown, then $Y(s_0)$ can be expressed as the linear function $Y(s_0) = \sum_{i=1}^{n} \lambda_i Y(s_i)$ and the constraint $\sum_{i=1}^{n} \lambda_i = 1$ is imposed to force $\hat{Y}(s_0)$ to be unbiased and the $\boldsymbol{\lambda}$ which minimizes the MSPE becomes

$$\boldsymbol{\lambda}^T = \left(\sigma + \mathbf{1}\frac{\mathbf{1} - \mathbf{1}^T\Sigma^{-1}\sigma}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\right)\Sigma^{-1} \quad \text{with} \quad \sigma^2(s_0) = C(1) - \sigma^T\Sigma^{-1}\sigma + \frac{\mathbf{1} - \mathbf{1}^T\Sigma^{-1}\sigma}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}$$

and $\hat{Y}(S_0) = \hat{\boldsymbol{\lambda}} Y(\boldsymbol{s})$ is called the ordinary kriging predictor. If further $\mu(s) = X\beta$, as in (1.0.1), we can construct the universal kriging predictor $\hat{Y}(S_0)$. For more details on universal kriging see Schabenberger and Gotway (2004). In general, the kriging predictor $\hat{Y}(s_0)$ is unbiased and the best linear predictor among all unbiased predictors. If we assume $Y(s)$ is a Gaussian random field, then $\hat{Y}(s_0)$ is also the best predictor among all linear or non-linear predictors. The kriging predictor also performs exact interpolation meaning $\hat{Y}(s_i) = Y(s_i)$ at all observed locations $s_i, i = 1, 2, ..., n$ if $Y(s_i)$ is used in prediction.

To evaluate the kriging predictions, MSPE and Log Score are commonly used, expressed as:

$$\text{MSPE} = \frac{1}{n} \sum_{i=1}^{n} \left\{ Y(\boldsymbol{s}_i) - \hat{Y}(\boldsymbol{s}_i) \right\}^2, \tag{1.1.5}$$

$$\text{Log Score} = \frac{1}{n} \sum_{i=1}^{n} \left[ \log(2\pi\hat{\sigma}_i^2) + \frac{(Y(\boldsymbol{s}_i) - \hat{Y}(\boldsymbol{s}_i))^2}{\hat{\sigma}_i^2} \right]. \tag{1.1.6}$$

where $\hat{\sigma}_i^2$ is the prediction variance for $Y(\boldsymbol{s}_i)$ give by (1.1.4).

### 1.1.1 Space-Time Data

Unlike in traditional statistics, spatial data does not typically have replicates unless we have multiple spatial samples observed over time. If we treat these spatial replicates as true independent replicates, we assume there is no temporal correlation in the data. Ignoring information about the behavior of the data across time is often an inappropriate assumption and can lead to obvious model misspecification. For example, many environmental data such as temperature and air pollutant concentrations are temporally correlated in addition to spatially correlated. For such data, we consider them as a spatio-temporal random field $\{Y(s,t),\ s \in \mathbb{R}^d,\ t \in \mathbb{N}\}$. Similarly

to the strictly spatial case, it is typical to model $Y(s,t)$ as:

$$Y(s,t) = \mu(s,t) + \nu(s,t),$$

where $\mu(s,t)$ can vary in space, time, or both and $\nu(s,t) \overset{iid}{\sim} (0, \Sigma(\theta))$. Since $\nu(s,t)$ is now spatially and temporally dependent, $\Sigma_{i,j}$ is expressed as $C(Y(s_i,t_i), Y(s_j,t_j))$. Again, most traditional models for $Y(s,t)$ assume second-order stationarity. $Y(s,t)$ is an SoS process if it has a constant mean and variance, i.e. $E[Y(s,t)] = \mu$, $Var[Y(s_t)] = \sigma^2$ and the covariance function can be expressed as:

$$Cov[Y(s,t), Y(s+h, t+u)] = C(h,u),\ (h,u) \in \mathbb{R}^d \times \mathbb{R},$$

where $h$ and $u$ are the spatial and temporal distances respectively between two observations. $C(h,0)$ is called the purely spatial covariance function and $C(0,u)$ the purely temporal covariance function. The process is spatially isotropic if $h = ||s_i - s_j||$. Just as with purely spatial covariance functions, a valid function $C(h,u)$ must be positive definite and given the process is SoS, we can define the spatio-temporal semivariogram as $\gamma(h,u) = C(0,0) - C(h,u)$. It is also of interest with space-time data to estimate $\Sigma(\theta)$ and make predictions at unknown locations and/or time points. It is important to note that the additional temporal dimension significantly increases the dimensions of $\Sigma$ which then decreases the feasibility of estimating such a function using likelihood estimation as the sample size gets larger. For this reason, a sample size of 10,000 observations (100 spatial locations and 100 time points) is considered a large spatio-temporal sample.

The simplest and most computationally feasible covariance models to fit are separable models, i.e. models we can express as $C(h,u) = C(0,u) \cdot C(h,0)$. For example,

below is a spatially isotropic and separable covariance model with exponential covariance function for both space and time:

$$C(h,u) = \sigma^2 e^{-\phi_t u} \cdot e^{-\phi_s h},$$

where $\phi_t$ and $\phi_s$ are the temporal and spatial range parameters respectively. A separable model allows us to write $\Sigma(\theta) = \sigma^2 V_t(\theta_t) \otimes V_s(\theta_s)$, and the profile log likelihood function from (1.1.3) can be written as:

$$l(\theta) = \frac{n_t \times n_s}{2} log\big(\frac{1}{n_t \times n_s}(Y - X\hat{\beta})^T V_t^{-1}(\theta_t) \otimes V_s^{-1}(\theta_s)(Y - X\hat{\beta})\big) + \frac{n_t}{2} log(|V_s(\theta_s)|) + \frac{n_s}{2} log(|V_t(\theta_t)|),$$

where $V_t(\theta_t)$ and $V_s(\theta_s)$ are the temporal and spatial correlation functions and $n_t$ and $n_s$ are the number of time points and spatial locations of the sample, respectively. The Kronecker product makes the maximum likelihood estimation of $\theta$ much more feasible. Although being convenient, the separability is not always an appropriate assumption. A general class of nonseparable isotropic covariance functions was proposed by Gneiting (2002):

$$C(h,u) = \frac{\sigma^2}{\psi(|u|^2)^{d/2}} \phi\left(\frac{||h||^2}{\psi(|u|)^2}\right), \ (h,u) \in \mathbb{R}^d \times \mathbb{R} \tag{1.1.7}$$

where $\phi(t)$, $t \geq 0$ is a complete monotone function and $\psi(t)$, $t \geq 0$ is a positive function with completely monotone derivatives. A specific example of a covariance model of this class can be seen in chapter 2.2.2.

## 1.2 Lattice Data

When the domain $D$ of $Y(s)$ is fixed and countable, the spatial data $\{Y(s),\ s \in D\}$

is known as lattice data, or areal data. The defining property of this type of data is that the locations are not random points but a finite number of areal regions often referred to as sites. An observation $Y(s_i)$ is typically aggregated data or summary data for a region $s_i$. In practice, these regions are irregularly shaped and unequally spaced rather than equally dispersed over a gridded area. For example, US census data are attribute summaries, such as population size, collected for each census tract. Due to patient confidentiality, epidemiology data can also be lattice data as it is often not released for exact geolocations but instead as summary data by county or zip-code. For example, the HIV rate, given by number of cases per 100,000 for each county in Illinois in 2012 (see Fig. 1.6). The domain $D$ is clearly fixed and countable as there are a finite number of counties in Illinois. The attribute of interest, HIV rate, is an aggregated value as it is a ratio of two values summed over each county. Because of privacy constraints, data from counties with very few cases or with small population sizes is suppressed by the counties (seen as negative values in Fig. 1.6).

A primary focus of lattice data is in modeling the spatial relationships in the data rather than prediction. A secondary goal in modeling lattice data is often identifying spatial risk factors associated with the attribute of interest. For example, the presence of other sexually transmitted diseases (STD) in neighboring counties in Illinois may contribute to a higher rate of HIV in a given county. Then it might be of interest to correlate the map shown in Fig. 1.6 to the one that shows the prevalence of other STD cases by county in Illinois during 2012. This can be achieved by appropriately modeling the relationship between the HIV rates and rates of other STDs. When modeling lattice data, it is usually appropriate to assume that neighboring regions are most similar to each other, that is there is a positive autocorrelation among regions. Thus, it is useful to construct a distance metric to describe the spatial relationship between regions. Since $D$ is a collection of finite locations and $s_i \in D$

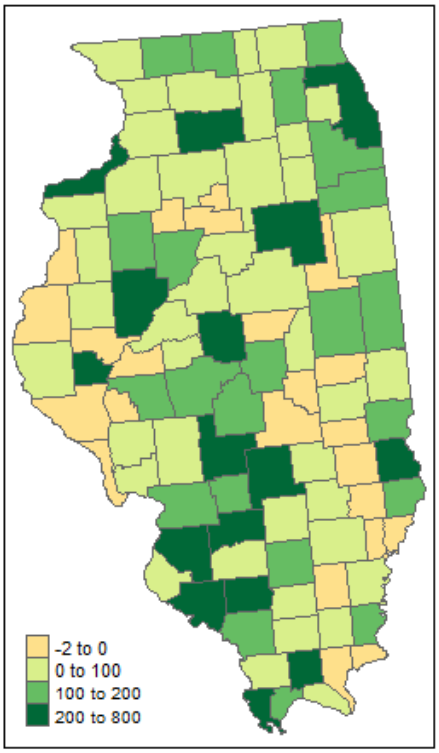


Figure 1.6: Rate in cases per 100,000 of HIV in Illinois in 2012 by county. (A negative number of cases indicates missing data due to privacy issues.)

denotes a spatial area rather than a specific coordinate, it is not so simple to compute the spatial "distance" $h_{i,j}$ between regions $s_i$ and $s_j$. One commonly used measure is the euclidean distance between similar points of reference within each spatial region, such as the centroid of each region. Defining the distances this way allows us to treat the data as geostatistical data and fit traditional covariance models as described in Section 1.1; although when modeling lattice data, there are no possible realizations between regions as there are in geostatistical data, e.g. there are no existing counties between bordering counties. An often more appropriate way to define distance in this case is to define a neighborhood matrix $B$ where, for example, $b_{i,j} = 1$ if regions $s_i$ and $s_j$ share a border and 0 otherwise. Simultaneous autoregressive (SAR) models and conditional autoregressive (CAR) models use neighborhood matrices to model spatial

dependence between regions. These models are special types of Markov random fields and are most popular for modeling lattice data.

Suppose that $\{Y(s_i) \in \mathbb{R},\ i = 1, 2, ..., n\}$ is a sample from a Gaussian spatial process observed at regions $s_i,\ i = 1, 2, ..., n$. Define B as the spatial dependence neighborhood matrix such that if $Y(s_i)$ depends positively on $Y(s_j)$, then the $i, j^{th}$ element of B, $b_{i,j} > 0$, otherwise $b_{i,j} = 0$. A commonly used structure for $B$ is the $0 - 1$ neighborhood matrix described previously. A SAR model is then defined as follows:

$$Y(s_i) = \mu_i + \sum_{j=1}^{n} b_{i,j}(Y(s_j) - \mu_j) + \epsilon_i,\ i = 1, 2, ..., n$$

where $\boldsymbol{\epsilon} \sim N(0, \Lambda)$ with diagonal variance matrix $\Lambda$. It can be shown trivially that the distribution of $Y(\boldsymbol{s})$ is jointly Gaussian:

$$\boldsymbol{Y} \sim MVN(\boldsymbol{\mu}, (I - B)^{-1}\Lambda(I - B^T)^{-1}).$$

A more widely used and more intuitive approach to modeling $Y(s_i)$ and its relationship with its neighbors is the CAR model which models $Y(s_i)$ conditioning on its neighbors, i.e. $Y_i|Y_{-i}$. Define the neighborhood matrix $C = [c_{i,j}]$, for which $c_{i,i} = 0$ and $c_{i,j} > 0$ if there is a positive pairwise dependence between regions $i$ and $j$, otherwise $c_{i,j} = 0$. If we assume $c_{i,j}\tau_j^2 = c_{j,i}\tau_i^2$, the CAR model is then expressed as

$$Y_i|Y_{-i} = \mu_i + \sum_{j=1}^{n} c_{i,j}(Y(s_j) - \mu_j) + v_i,\ i = 1, 2, ...n \tag{1.2.1}$$

$$Var[Y_i|Y_{-i}] = \tau_i^2$$

and the joint distribution can be described as

$$Y(\boldsymbol{s}) \sim MVN(\boldsymbol{\mu}, (I - C)^{-1}M)$$

where $M = \text{diag}(\tau_1^2, \tau_2^2, ..., \tau_n^2)$. The most notable difference between the CAR and SAR models in the Gaussian case is that $v_i$ is uncorrelated with $Y(s_j), j \neq i$ while $\epsilon_i$ is not, which is why the former is preferred for estimation and interpretation of model parameters. It is shown in Cressie (1991), that any Gaussian SAR model can be represented as a Gaussian CAR and that the two models are equivalent when their variance matrices are equal, e.g $(I - C)^{-1}M = (I - B)^{-1}\Lambda(I - B^T)^{-1}$.

SAR and CAR models often seek to model the probability distribution of the process $\{Y(s),\ s \in D\}$ and are commonly seen in a Bayesian hierarchical framework, in particular when modeling disease incidence as in Section 3. Suppose we observe the disease incidences in $Y(s_i), ..., Y(s_n)$ in regions $s_i, i = 1, 2, ..., n$ with corresponding population at risk $N_i$ and underlying disease rate $p_i$. An expected number of cases in a given region, $E_i$, that serves as a baseline measure in order to identify regions with elevated risks, is determined by population size and demographic of each county. A region with elevated risk is indicated by $p_i > 1$. If a disease is considered rare or the region's population $N_i$ is small, it is likely that $E_i$ is small. In practice, Bayesian hierarchical models are used widely for estimating both a fixed set of covariates which contribute to the disease risk and a set of spatial random effects that describe the spatial variability of disease. In a three-level Bayesian hierarchical spatial model for disease incidence, the fixed covariates are modeled in the first level, spatial correlation is modeled in the second level via random effects, and priors for the hyperparameters such as $\mu$, $\boldsymbol{\beta}$, and $\tau_i^2$ are specified in the third level. A general framework for this

model is given by Lee (2011) as below:

$$Level\ I: \quad Y_k|E_i, p_i \sim Poisson(E_i p_i),\ i = 1, ..., n$$

$$Level\ II: \ log(p_i) = \mu + x_i^T \boldsymbol{\beta} + \phi_i$$

$$Level\ III(priors): \ \boldsymbol{\beta} \sim MVN(0, \sigma_\beta^2 I_p), \mu \sim N(0, \sigma_\mu^2)$$

A CAR model is commonly used as a prior distribution for the spatial random effect $\phi_i$ by modeling $\phi_i|\phi_{-i}$, where $-i$ denotes all neighboring regions of $s_i$. If we define the neighborhood matrix $B$ as above, the simplest CAR prior, proposed by Besag et al. (2011) called the intrinsic autoregressive model (IAR), is a special case of (1.2.1) and is given by

$$\phi_i|\phi_{-i}, \tau_i^2 \sim N\left(\frac{1}{n_i}\sum_{j\sim i}\phi_j, \frac{\tau_i^2}{n_i}\right) \tag{1.2.2}$$

where $n_i$ is the number of neighbors for region $s_i$ and $j \sim i$ indicates that region $s_j$ is a neighbor of $s_i$. This form can be derived from (1.2.1) if we define $C$ to be the $0-1$ neighborhood matrix. Inference can be made for this model using techniques based on Markov chain Monte-Carlo (MCMC) simulations. While $\mu$ and $\boldsymbol{\beta}$ are commonly assigned Gaussian priors, the variance parameter, $\tau^2$, is assigned an inverse gamma prior or a uniform prior favored by Gelman (2006) as this prior tends to be less sensitive to initial parameter values when the true variance $\tau^2$ is near zero. A Poisson link function is typical in disease mapping since many diseases of interest, such as cancer and HIV, are considered rare. In addition to accounting for spatial correlation, the random effects $\boldsymbol{\phi}$ are also useful for correcting the overdispersion that can occur due to the property of $E[Y(s_i)] = Var[Y(s_i]$ for a Poisson process. In most disease studies, $E[Y(s_i)] < Var[Y(s_i)]$. Lee (2011) points out that the existence of overdispersion and spatial correlation is most likely due to unknown risk factors.

# Chapter 2

# Dimension Expansion for Nonstationary Space-Time Data

## 2.1 Background

Interests in spatio-temporal data analysis are rising as such data has become more abundant in agriculture, atmospheric, hydrological and other environmental sciences. For example, the amount of precipitation at any given hour and unmonitored location depends on the precipitation in the surrounding areas prior to that hour. Understanding the underlying structures of spatio-temporal processes largely relies on covariance and variogram modeling techniques, the majority of which assume for simplicity that the process is stationary. This assumption, however, can be violated as real environmental spatio-temporal processes are very often inherently non-stationary in both space and time. For example, precipitation measures can greatly vary if taken prior to or during a moving storm as well as at locations outside the edges or at the eye of the storm.

The nonstationarity can be observed in either or both the space and time domains. Let $Y(\boldsymbol{s}, t)$ denote a space-time random field, where $(\boldsymbol{s}, t) \in \mathbb{R}^d \times \mathbb{R}$. The process $Y(\boldsymbol{s}, t)$ is nonstationary if either $E\{Y(\boldsymbol{s}, t)\}$ varies or the covariance $C\{Y(\boldsymbol{s}_i, t_i), Y(\boldsymbol{s}_j, t_j)\}$ for $(\boldsymbol{s}_i, t_i)$ and $(\boldsymbol{s}_j, t_j)$ in $\mathbb{R}^d \times \mathbb{R}$ depends on locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ and/or time points $t_i$ and $t_j$, rather than only their space-time lag. We focus on the nonstationarity that rises only from the heterogeneous correlation between observations separated by the same space-time lags, assuming constant mean and variance. This assumption is reasonable as large-scale spatially and temporally varying means and variances are often removed before analyzing the dependence structure of $Y(\boldsymbol{s}, t)$. Nonstationarity can be visually diagnosed by examining the variability of the empirical covariances at the same space-time lags. Large variability indicates nonstationarity. A more rigorous test for stationarity is seen in Jun and Genton (2012). Another indication of nonstationarity can be derived from the physical formulation of the process. For instance, if a random process such as precipitation is generated from complex climate dynamics and affected by many environmental factors, then it is most likely nonstationary.

Modeling nonstationarity has been investigated for years and has recently received more attention due to high demands in practice. Most available models focus only on the nonstationarity in the spatial domain. For example, Sampson and Guttorp (1992) and Guttorp et al. (1994) proposed a nonstationary modeling framework based on deformation techniques (see also Schmidt and O'Hagan, 2003; Aberg et al., 2005). Higdon (1998) and Higdon et al. (1999) developed a flexible nonstationary model through the convolution of stationary spatial processes with spatially varying kernels (see also Risser and Calder, 2015; Wikle, 2002). Fuentes (2001, 2002) reframed the convolution method by allowing the latent process to be spatially dependent and Paciorek and Schervish (2004) generalized the kernel convolution approach to a class of nonstationary covariance functions. Bornn et al. (2012) proposed a different idea

to model nonstationarity which assumed the nonstationary process is a projection of a stationary process in higher dimensions.

Recently, with the onset of big data, the spatial random effects model (e.g. Cressie and Johannesson, 2008; Katzfuss, 2013; Nychka et al., 2015) and the predictive process model (e.g. Banerjee et al., 2008; Eidsvik et al., 2012; Ren and Banerjee, 2013) were developed to make the computation feasible while carrying the ability to model nonstationarity. The main idea of those two types of models is to represent the full-rank random processes using reduced-rank basis functions or predictive processes. Although these models show big advantages in capturing large-scale correlation, Stein (2014) demonstrated possible statistical inefficiency of low rank models for spatial interpolation. The class of Markov random field models (Bolin and Lingren, 2011) constructed from nested statistical partial differential equations serves as another computationally efficient method to capture nonstationary features in spatial data.

Compared to the vast literature for modeling nonstationarity in space, nonstationarity in both space and time is studied much less extensively. Ma (2002) derived nonstationary space-time covariance functions by applying certain kernels to stationary covariance functions. Garg et al. (2012) extended the idea of convolution processes in Higdon et al. (1999) and Paciorek and Schervish (2004) to model a nonstationary spatio-temporal Gaussian process. Set in a Bayesian framework, Sigrist et al. (2012) proposed a dynamic nonstationary spatio-temporal model for short term precipitation by linking the advection parameter, describing the horizontal transport of rainfall, of the convolution kernel to an external wind vector, measured independently of precipitation. Stroud et al. (2001) achieved nonstationarity in time by allowing the regression coefficients in a locally weighted mixture model to temporally vary. Huang and Hsu (2004) extended Wikle and Cressie (1999) to develop a space-time Kalman filter where the spatio-temporal covariance function depends on covariates.

We propose modeling nonstationarity in both space and time for a spatio-temporal process by applying the dimension expansion technique described in Bornn et al. (2012). Spatio-temporal data often exhibit temporal correlation in addition to spatial and both may not be stationary. It is thus important to build a flexible space-time model which accommodates nonstationarity in both domains. Unlike previously proposed nonstationary space-time models, dimension expansion allows us to make use of the plethora of stationary covariance models available. Numerous studies have been conducted on developing stationary space-time covariance models, e.g., Gneiting (2002), Stein (2005), Fonesca and Steel (2011) and Choi et al. (2013). Our model is demonstrated to have advantages over the nonstationary spatial model in Bornn et al. (2012) as well as over a combination of this nonstationary spatial model with a stationary temporal model. Compared to the low rank models, our model focuses on fine-scale correlation.

This chapter is organized as follows: Section 2.2 details the dimension expansion technique, investigates model fitting and proposes an approach to deal with computation for large datasets. Section 2.3 studies the properties of the nonstationary model and compares it to other models through simulated data, while Section 2.4 applies the nonstationary model to spatio-temporal windflow and streamflow datasets. Finally, Section 2.5 provides a conclusion and discussion.

## 2.2 Dimension Expansion in Space and Time

Perrin and Meiring (2003) showed that any nonstationary random field in $\mathbb{R}^n$, with moments of at least order 2, can be interpreted as a lower-dimensional representation of a second-order stationary random field in a higher dimensional space $\mathbb{R}^{2n}$, provided that the components of the random vector have identical expectations

and variances. Further they established that a bijective mapping exists between the two spaces. Perrin and Schlather (2007) generalized this idea and proved that any Gaussian random vector can be interpreted as a sample from a stationary random function on a graph in $\mathbb{R}^d$, $d \geq 2$, subject to the same moment constraints as in Perrin and Meiring (2003). An intuitive example is that of suppressing the elevation from a stationary 3-$d$ temperature process, leading to a nonstationary 2-$d$ process.

Based on the above theoretical results, Bornn et al. (2012) justified their dimension augmentation procedure by stating that a realization of a nonstationary Gaussian process in $\mathbb{R}^d$ may be interpreted as a realization of a stationary field in $\mathbb{R}^{d+p}$ for $p > 0$ under appropriate moment constraints. Since a spatio-temporal random field can be considered a random field in $\mathbb{R}^d \times \mathbb{R}$, it follows that under appropriate moments constraints, the nonstationary space-time process $Y(\boldsymbol{s}, t)$ with $(\boldsymbol{s}, t) \in \mathbb{R}^d \times \mathbb{R}$ can be represented by a stationary process $Y([\boldsymbol{s}, \boldsymbol{z}], [t, \boldsymbol{w}])$ with $[\boldsymbol{z}, \boldsymbol{w}] \in \mathbb{R}^p \times \mathbb{R}^q$ for $p + q > 0$, that is, the process $Y$ is stationary in the space $\mathbb{R}^{d+p} \times \mathbb{R}^{1+q}$ although nonstationary in $\mathbb{R}^d \times \mathbb{R}$. Then, after estimating the augmented dimensions $\boldsymbol{z}$ and $\boldsymbol{w}$, any valid stationary space-time covariance model in $\mathbb{R}^{d+p} \times \mathbb{R}^{1+q}$ will be applicable to $Y([\boldsymbol{s}, \boldsymbol{z}], [t, \boldsymbol{w}])$. Following traditional spatio-temporal modeling, we treat the time and spatial domains separately to reflect their different characteristics.

### 2.2.1 Review of dimension expansion in space

Let $Y(\boldsymbol{s})$, $\boldsymbol{s} \in \mathbb{R}^d$ be a nonstationary spatial process, and suppose $\boldsymbol{z} \in \mathbb{R}^p, p > 0$ are the latent $p$ dimensions such that $Y([\boldsymbol{s}, \boldsymbol{z}])$ is stationary with semivariogram $\gamma_\theta(\cdot)$ defined as

$$\gamma_\theta([\boldsymbol{s}_i, \boldsymbol{z}_i] - [\boldsymbol{s}_j, \boldsymbol{z}_j]) = \frac{1}{2} E[\{Y([\boldsymbol{s}_i, \boldsymbol{z}_i]) - Y([\boldsymbol{s}_j, \boldsymbol{z}_j])\}^2],$$

where $[\boldsymbol{s}_i, \boldsymbol{z}_i]$ is the $i$th element of $[\boldsymbol{S}, \boldsymbol{Z}]$, the concatenation of the dimensions $\boldsymbol{S}$ and

$\boldsymbol{Z}$. One can model the nonstationarity in $Y(\boldsymbol{s})$ by a stationary variogram model defined for $Y[\boldsymbol{s}, \boldsymbol{z}]$. Note any valid stationary variogram model is a legitimate choice for $\gamma(\cdot)$. Bornn et al. (2012) proposed to estimate the latent dimensions $\boldsymbol{z}$ using lasso-penalized least squares:

$$\widehat{\theta}, \boldsymbol{z} = \text{argmin}_{\theta,\boldsymbol{z}} \sum_{i<j} [\nu_{i,j} - \gamma_\theta\{d_{i,j}([\boldsymbol{S}, \boldsymbol{Z}])\}]^2 + \lambda_{s,1} \sum_{k=1}^{p} ||\boldsymbol{Z}_{\cdot k}||_1, \tag{2.2.1}$$

where $\nu_{i,j}$ are the moment estimates of $\gamma_\theta(\cdot)$, $d_{i,j}([\boldsymbol{S}, \boldsymbol{Z}])$ is the $i, j$th element of the distance matrix of the augmented locations $[\boldsymbol{S}, \boldsymbol{Z}]$, $\boldsymbol{Z}_{\cdot k}$ is the $k$th column (dimension) of $\boldsymbol{Z}$ and $||\cdot||_1$ is the $L_1$ norm. In their model fitting, time points at different locations are treated as replicates to compute the empirical spatial variogram. The purpose of the lasso group penalty parameter, $\lambda_{s,1}$, is to regularize the estimation of $\boldsymbol{z}$ and prevent an overfitting of the spatial dimensions by controlling the dimension sparsity of $\boldsymbol{z}$. Finally, a bijective function $f(\cdot)$ between $\boldsymbol{s}$ and $\boldsymbol{z}$ can be established using thin-plate splines based on the estimates of $\boldsymbol{z}$.

### 2.2.2 Extend dimension expansion to time

In addition to spatial correlation, spatio-temporal data often exhibit correlation in time, which can also be nonstationary. Nevertheless, it is still quite common that nonstationary spatio-temporal models only address spatial nonstationarity. Allowing dimension expansion in both the space and time domains, however, allows us to relax all stationary assumptions while enjoying already established stationary models. We now illustrate how to enable the rich class of stationary models to capture nonstationary features in data.

We will use two classes of stationary space-time covariance models to exemplify this approach. These two classes of models will also be used in our simulation studies

in Section 2.3. The first one is the space-time separable covariance function. The separable model is often an oversimplification of the dependence structure in real data, yet it is widely used for its simplicity and computational efficiency. Without loss of generality, we choose the exponential covariance function for both the space and time domain:

$$C(h,u) = \sigma^2 \exp(-\phi_s h) \exp(-\phi_t u), \tag{2.2.2}$$

where $h$ and $u$ are the Euclidian distances of the spatial lag $\boldsymbol{h} \in \mathbb{R}^{d+p}$ and temporal lag $\boldsymbol{u} \in \mathbb{R}^{1+q}$, respectively, and $\phi_s$ and $\phi_t$ are range parameters in space and time domains. Suppose $([\boldsymbol{s}_i, \boldsymbol{z}_i], [t_i, \boldsymbol{w}_i])$ and $([\boldsymbol{s}_j, \boldsymbol{z}_j], [t_j, \boldsymbol{w}_j])$ are two locations in the expanded space, then

$$\boldsymbol{h} = [\boldsymbol{s}_i - \boldsymbol{s}_j, \boldsymbol{z}_i - \boldsymbol{z}_j],\ \boldsymbol{u} = [t_i - t_j, \boldsymbol{w}_i - \boldsymbol{w}_j],$$

and $h = ||\boldsymbol{h}||,\ u = ||\boldsymbol{u}||$, where $||\boldsymbol{a}||$ is the Euclidean norm of a vector $\boldsymbol{a}$. Although model (2.2.2) is obviously a stationary model in $\mathbb{R}^{d+p} \times \mathbb{R}^{1+q}$, it can model the nonstationary correlation structure of the process $Y$ in a lower dimensional space, e.g., in $\mathbb{R}^d \times \mathbb{R}$.

The second example is the class of nonseparable space-time models. These models are more flexible though usually involve more parameters and are more computationally challenging. Gneiting (2002) proposed a general form for nonseparable covariance functions, and we adopted one particular form for our illustration:

$$C(h,u) = \begin{cases} \sigma^2(a|u|^{2\alpha}+1)^{-1}, & \text{if } h = 0 \\ \sigma^2(a|u|^{2\alpha}+1)^{-1} \exp\left\{\frac{-ch}{(a|u|^{2\alpha}+1)^{\beta/2}}\right\}, & \text{otherwise,} \end{cases} \tag{2.2.3}$$

where $h$ and $u$ follow the notation in model (2.2.2), $a$ and $c$ are nonnegative scaling parameters of time and space respectively, $\alpha \in (0, 1]$ is a smoothing paramter and $\sigma^2$ is the variance of the process. The parameter $\beta \in [0, 1]$ measures the space-time interaction where a larger $\beta$ indicates a stronger interaction or weaker separability between space and time components. Model (2.2.3) is developed for stationary processes, but with distance lags $\boldsymbol{u} \in \mathbb{R}^{1+q}$ and $\boldsymbol{h} \in \mathbb{R}^{d+p}$, it is capable of modeling nonstationary processes observed in lower dimensional spaces. Although dimension expansion applies to all covariance structures, separable or nonseparable, a separability test, e.g. Li et al. (2007), can help choose an appropriate model.

### 2.2.3 Learning latent dimensions and model fitting

Let $\boldsymbol{\theta}$ denote the vector of parameters in the stationary space-time covariance function. The nonstationary model using expanded dimension(s) has three sets of unknown parameters: $\boldsymbol{\theta}$, $\boldsymbol{z}$, and $\boldsymbol{w}$. To reflect that $\boldsymbol{z}$ and $\boldsymbol{w}$ essentially form two continuous processes, we denote them as $\boldsymbol{z}(\boldsymbol{s})$ and $\boldsymbol{w}(t)$. Note that the dimensions of $\boldsymbol{z}(\boldsymbol{s})$ and $\boldsymbol{w}(t)$, $p$ and $q$, also remain unknown. A simultaneous estimation for all parameters at once is obviously not ideal from a numerical point of view. We therefore propose to estimate the parameters in two steps. First we estimate $\boldsymbol{z}(\boldsymbol{s})$ and $\boldsymbol{w}(t)$ using penalized least squares and then we estimate $\boldsymbol{\theta}$ using likelihood method based on the estimated latent dimensions.

To estimate $\boldsymbol{z}(\boldsymbol{s})$ and $\boldsymbol{w}(t)$, it seems natural to extend the estimation method in (2.2.1) into the space-time context so that we have

$$\begin{aligned} \widehat{\boldsymbol{\theta}}, \boldsymbol{z}(\boldsymbol{s}), \boldsymbol{w}(t) = \; & \text{argmin}_{\theta,w} \textstyle\sum_{i<j} \big[\widehat{C}\{(\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j)\} \\ & -C_{\boldsymbol{\theta}}(||\boldsymbol{s}_i - \boldsymbol{s}_j, \boldsymbol{z}(\boldsymbol{s}_i) - \boldsymbol{z}(\boldsymbol{s}_j)||, ||t_i - t_j, \boldsymbol{w}(t_i) - \boldsymbol{w}(t_j)||)\big]^2 \\ & +\lambda_{s,1} \textstyle\sum_{k=1}^{p} ||\boldsymbol{Z}_{.k}||_1 + \lambda_{t,1} \sum_{k=1}^{q} ||\boldsymbol{W}_{.k}||_1, \end{aligned} \tag{2.2.4}$$

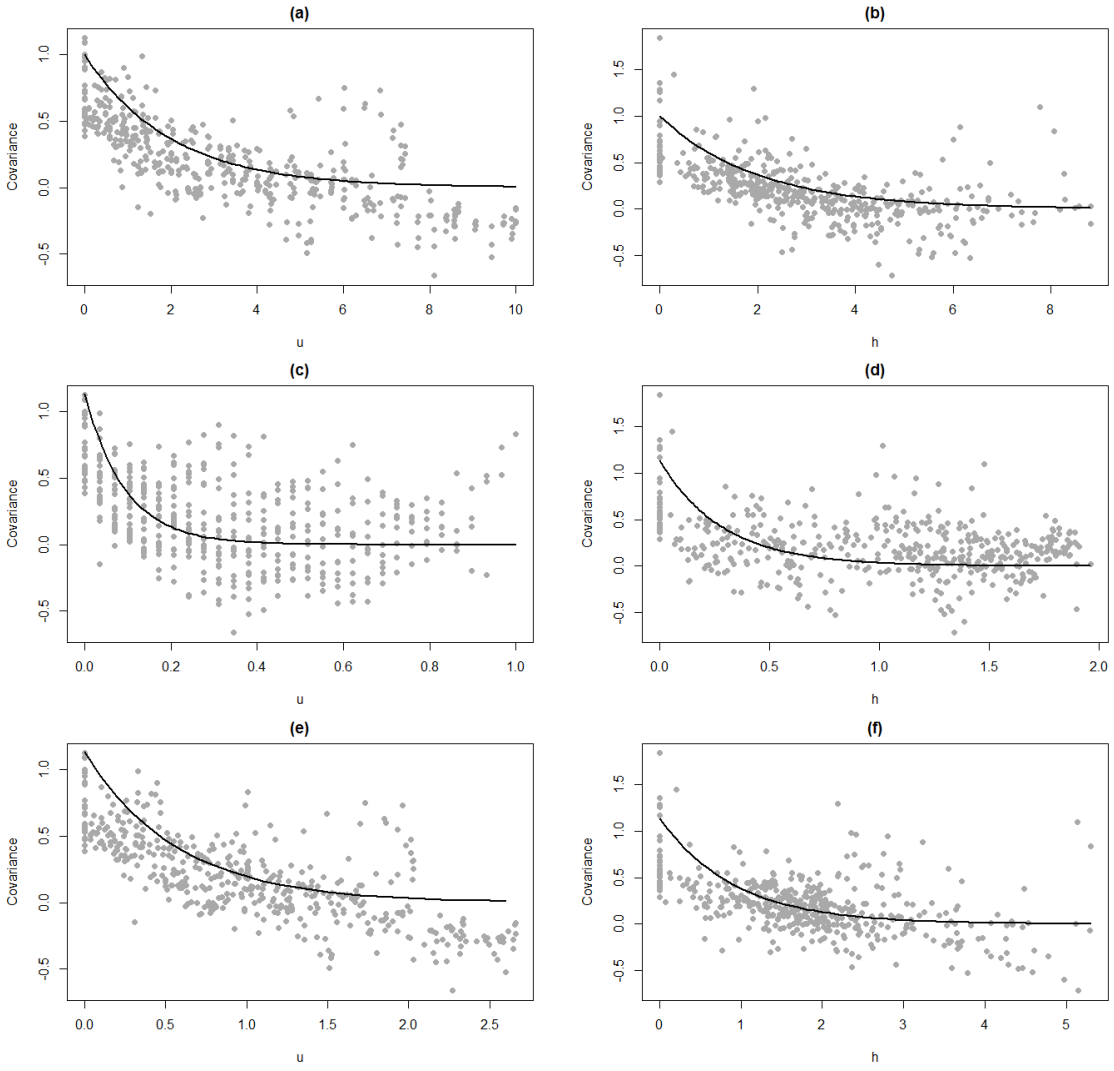


Figure 2.1: Empirical temporal and spatial covariance plots with (a)-(b): true dimensions; (c)-(d): reduced dimensions; and (e)-(f): the expanded dimensions using the estimated latent dimensions. The lines are the fitted exponential model using the maximum likelihood method.

where $\widehat{C}(\cdot)$ is an empirical estimate of the covariance function. However, estimating $\boldsymbol{z}(\boldsymbol{s})$ and $\boldsymbol{w}(t)$ simultaneously is not accessible in our case as there are simply no replicates to compute $\widehat{C}(\cdot)$ unless there are repeat measurements at each $(\boldsymbol{s}_i, t_i)$. To make the estimation procedure more feasible, we propose to estimate $\boldsymbol{z}(\boldsymbol{s})$ and $\boldsymbol{w}(t)$ separately.

Treating the spatial locations as replicates for a given time point, we first estimate the temporal covariance function using moment estimator:

$$\widehat{C}(t_i, t_j) = \frac{1}{|D|}\sum_{\boldsymbol{s}\in D}\{Y(\boldsymbol{s}, t_i)Y(\boldsymbol{s}, t_j)\},$$

where $|D|$ is the cardinality of the spatial domain $D$. Then we estimate $\boldsymbol{w}(t)$ by minimizing the following lasso penalized least squares:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_t, \boldsymbol{w} = \; & \text{argmin}_{\boldsymbol{\theta}_t, \boldsymbol{w}} \textstyle\sum_{i<j}[\widehat{C}(t_i, t_j) - C_{\boldsymbol{\theta}_t}\{||t_i - t_j, \boldsymbol{w}(t_i) - \boldsymbol{w}(t_j)||\}]^2 \\ & + \lambda_{t,1} \textstyle\sum_{k=1}^{q} ||\boldsymbol{w}_{.k}||_1,\end{aligned} \tag{2.2.5}$$

where $\boldsymbol{\theta}_t$ is the vector of parameters only involved in the pure temporal covariance function. The estimation of $\boldsymbol{z}(\boldsymbol{s})$ follows Bornn et al. (2012) by treating the observations at different time points as replicates for a given location. After we estimate the point estimates $\boldsymbol{z}(s_i)$ and $\boldsymbol{w}(t_j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, T$, we then build thin-plate splines $g_s(\boldsymbol{s}) \approx \boldsymbol{z}$ and $g_t(\boldsymbol{t}) \approx \boldsymbol{w}$, with smoothing parameters $\lambda_{s,2}$ and $\lambda_{t,2}$ respectively, to obtain the continuous processes of $\boldsymbol{z}(\boldsymbol{s})$ and $\boldsymbol{w}(t)$. Another example of using thin-plate splines in modeling nonstationary covariance structures can be seen in Sampson and Guttorp (1992). The choice of all tuning parameters is determined by cross validation for which the target is to minimize the prediction errors. Finally, we reestimate $\boldsymbol{\theta}$ using the maximum likelihood (ML) method based on the observed dimension and estimated latent dimensions $\{[\boldsymbol{s}, \boldsymbol{z}], [t, \boldsymbol{w}]\}$. Although estimates for $\boldsymbol{\theta}$

are obtained in (2.2.1) and (2.2.5), only the estimates for $\boldsymbol{w}$ and $\boldsymbol{z}$ are kept from these steps. ML estimates for $\boldsymbol{\theta}$ are preferred since the MLE is asymptotically most efficient.

The separate estimation of latent dimensions in space and time may seem suboptimal, especially when fitting a nonseparable covariance model. We carried out a small simulation estimating $\boldsymbol{w}$ and $\boldsymbol{z}$ together by replacing $\widehat{C}\{(\boldsymbol{s}_i, t_i), (\boldsymbol{s}_j, t_j)\}$ in (2.2.4) with $Y(\boldsymbol{s}_i, t_i)Y(\boldsymbol{s}_j, t_j)$ so that all interactions between space and time are taken into account. Due to the large number of unknown parameters to be estimated and great variability in each individual pair, the simultaneous estimates of $\boldsymbol{w}$ and $\boldsymbol{z}$ performed worse than our proposed method in addition to being overly time consuming. The deteriorated performance of estimators that attempt to use all individual pairs directly is also observed in Choi et al. (2013).

Additionally we explored two other methods to estimate $\boldsymbol{w}$ and $\boldsymbol{z}$: an iterative procedure, performed by estimating $\boldsymbol{w}$ given $\boldsymbol{z}$ and then $\boldsymbol{w}$ given $\boldsymbol{z}$ using (2.2.4) until convergence, and a 2-step hybrid procedure where we estimate $\boldsymbol{z}$ using (2.2.1) and then $\boldsymbol{w}$ given $\boldsymbol{z}$ using (2.2.4). Although we found both methods performed comparably to our current method, the relative computational efficiency makes the current method preferable. Finally, when data contain outliers, we can consider a more robust estimation method by replacing the classical empirical variogram and covariance estimates in (2.2.1) and (2.2.5) respectively with more robust estimates (e.g. Cressie and Hawkins, 1980). A simulation study not presented here shows that the robust estimator may improve the prediction when outliers are present.

### 2.2.4 Reducing computational burden for a large dataset

Dimension expansion can be computationally expensive as the number of spatial locations, $n$, or time points, $T$ increases. Bornn et al. (2012) mentioned that more

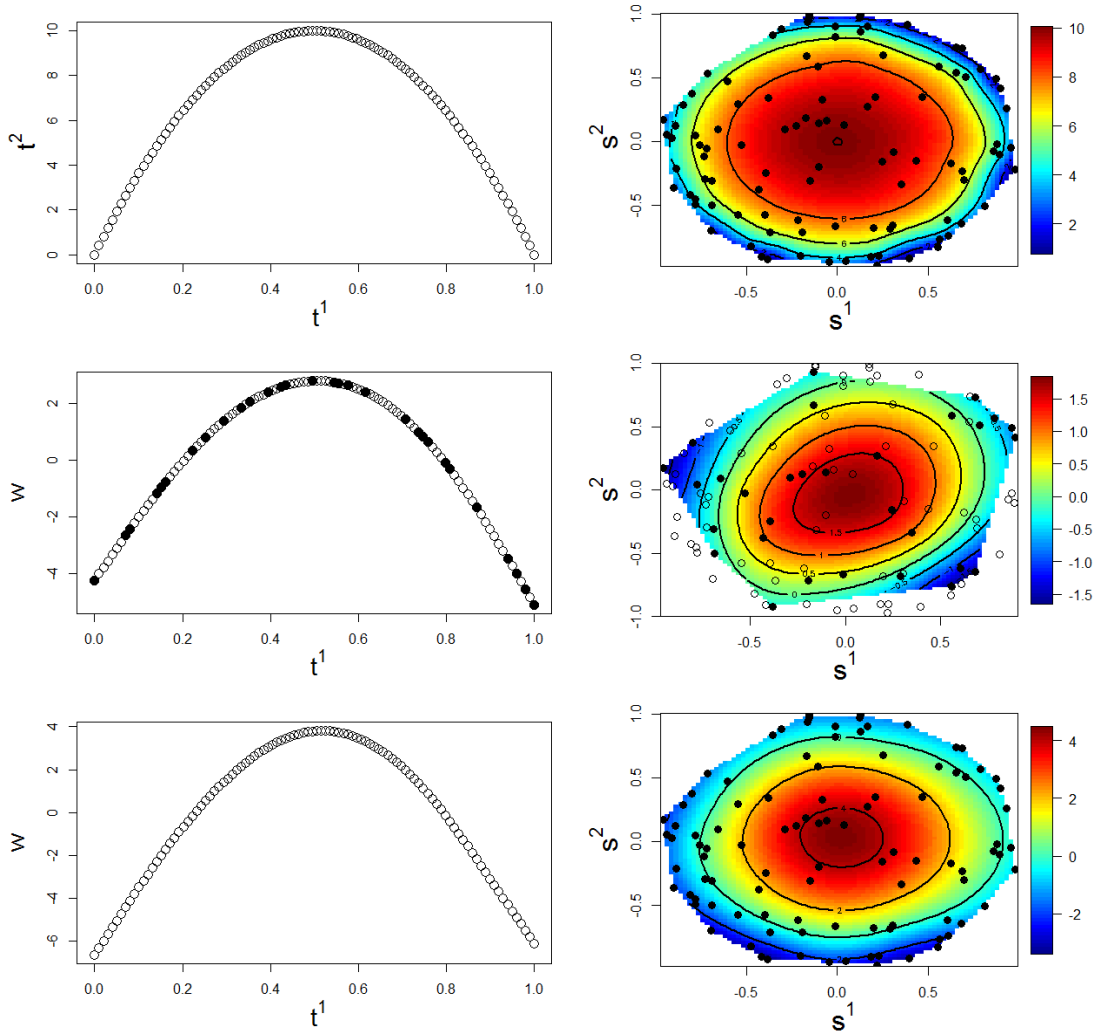


Figure 2.2: Upper: true latent dimensions in time (left panel) and space (right panel); Middle: estimated latent dimensions using 900 of observations. The solid points in the left (right) panel represent the sampled time (spatial) locations; Bottom: estimated latent dimensions using all observations.

complex optimization methods are necessary when $n$ exceeds 100 and the number of estimated dimensions, $p$ is greater than 3. To tackle the computational issue without resorting to complex optimization methods, we propose to use a small subset of spatial locations and time points to estimate $\boldsymbol{z}(\boldsymbol{s})$ and $\boldsymbol{w}(t)$ respectively. Specifically, we take a subset of $n_0$ locations and $T_0$ time points and then estimate $\boldsymbol{z}(s_i)$ and $\boldsymbol{w}(t_j)$ for $i = 1, \ldots, n_0$ and $j = 1, \ldots, T_0$ based on this subset. Then we build thin-plate splines $g_t(\cdot)$ and $g_s(\cdot)$, with respective smoothing parameters $\lambda_{s,2}$, and $\lambda_{t,2}$, only based on these point estimates and apply those functions to the whole domain to obtain the continuous processes $\boldsymbol{z}(\boldsymbol{s})$ and $\boldsymbol{w}(t)$. Tuning parameters $\lambda_{t,1}, \lambda_{s,1}, \lambda_{t,2}$, and $\lambda_{s,2}$ are chosen by minimizing the drop-one hundred MSPE of the $n_0 \times T_0$ sampled observations. The subset of spatial locations can be randomly sampled from the spatial domain, whereas we suggest the $T_0$ time points to be selected strategically by first sectioning the $T$ time points into blocks and then sampling a fraction of time points from each block. We recommend to always include the first and last time points in the sample to avoid extrapolation when using thin-plate splines to estimate $\boldsymbol{w}(t)$. In the case of a large number of spatial locations over a larger spatial domain, a similar sampling strategy may also need to be implemented to ensure that the selected sample has adequate spatial coverage. Note that in the case of fitting a separable space-time covariance function, one can take advantage of properties of the Kronecker product to further reduce the computation.

## 2.3 Simulation study

### 2.3.1 Basic setup

We perform simulation studies to demonstrate the advantages of using our method

when the data are nonstationary in both space and time. Two space-time covariance functions are used to simulate spatio-temporal processes. One is the separable model in (2.2.2) and the other is the nonseparable model in (2.2.3). For each model type, we set parameters at different values to eliminate the effect of particular parameters on the simulation results. Specifically, as shown in Table 2.1, we set $\phi_t = \phi_s = 0.3, 0.5, 0.8$ in the separable model (2.2.2) to simulate three processes representing strong, fair and weak spatial and temporal correlations, respectively. To simulate a variety of processes with different properties using the nonseparable model (2.2.3), we set smoothing parameter $\alpha = 0.5$ and set $\beta$ =0.5, 1 to represent weak and strong space-time interactions. We also set $a$ =2, 7 and $c$ =1, 2 to represent strong and weak temporal and spatial correlation, respectively. We set $\sigma^2 = 1$ for all separable and nonseparable models.

At each parameter setting, 200 zero-mean Gaussian spatio-temporal processes, $\{Y(\boldsymbol{s}_i, \boldsymbol{t}_j)\}$, $i = 1, \ldots, 30$, $j = 1, \ldots, 30$, are simulated at 30 randomly chosen locations and 30 time points. Here we set spatial locations $\boldsymbol{s}_i = (s_i^1, s_i^2, s_i^3) \in \mathbb{R}^3$ and time locations $\boldsymbol{t}_j = (t_j^1, t_j^2) \in \mathbb{R}^2$, that is, we generate stationary Gaussian random processes in $\mathbb{R}^3 \times \mathbb{R}^2$. The time points $t_j^1$ are equally spaced between 0 and 1 and the latent dimension $t_j^2$ are generated as a quadratic function of $t_j^1$. The spatial locations $(s_i^1, s_i^2)$ are generated uniformly on a unit circular region centered at (0, 0) and the latent dimension $s_i^3$ is generated such that $(s_i^1, s_i^2, s_i^3)$ are centered on a three-dimensional half-ellipsoid. Then we project the processes onto the reduced dimensional space $\mathbb{R}^2 \times \mathbb{R}$ by retaining only $(s_i^1, s_i^2)$ and $t_j^1$ in their spatial and temporal domains. The spatio-temporal processes in the $\mathbb{R}^2 \times \mathbb{R}$ space exhibit nonstationarity.

In the simulations, all parameters are estimated using (2.2.1) and (2.2.5). Note that in the cross validation for identifying the tuning parameters for the estimation of $\boldsymbol{z}(s)$ and $\boldsymbol{w}(t)$, we again propose to substitute the nonseparable covariance function by

its corresponding separable form to ease the computation. Since we will eventually discard the covariance parameter estimates yielded during the estimation of latent dimension and use ML to estimate $\boldsymbol{\theta}$ of the whole space-time covariance function based on the recovered latent dimensions, the approximations of $\boldsymbol{\theta}$ obtained when estimating latent dimensions have very little effect on covariance modeling. Through another small simulation study we found that the prediction errors using our method are comparable to those using the full nonseparable model to estimate the tuning parameters and latent dimensions.

To evaluate the strength of modeling nonstationarity in both space and time, we compare its predictive performance to several other models. The nonstationary separable model allows us to compare it with the (i) true model as a reference; (ii) stationary space-time model which simply applies a stationary model to the observed dimensions; (iii) Bornn et al.'s method that treats the temporal replicates as independent, hereinafter referred to as BSZ. This model takes advantage of the latent dimensions in the spatial domain but ignores the correlation in time; (iv) Bornn et al.'s method in combination with a stationary temporal model, hereinafter referred to as BSZ+ST. This model also considers the latent dimensions in the spatial domain but not in the time domain. The nonstationary nonseparable model only allows us to compare with models (i), (ii) and (iv). We use the mean squared prediction error (MSPE) and log scores of drop-one predictions as measures for the predictive performance:

$$\text{MSPE} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{j=1}^{T}\left\{Y(\boldsymbol{s}_i,t_j)-\hat{Y}(\boldsymbol{s}_i,t_j)\right\}^2,$$

$$\text{Log Score} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{j=1}^{T}\left[\log(2\pi\widehat{\sigma}^2_{ij})+\{Y(\boldsymbol{s}_i,t_j)-\hat{Y}(\boldsymbol{s}_i,t_j)\}^2/\widehat{\sigma}^2_{ij}\right]$$

where $\widehat{\sigma}^2_{ij}$ is the prediction variance for $Y(\boldsymbol{s}_i,t_j)$.

Figure 2.1 shows an example of the empirical covariance in the expanded dimensional space using the estimated latent dimensions compared to that in the reduced dimensional space and true dimensions, based on a randomly chosen simulated dataset using the separable model. The high variability of empirical covariances at each spatial lag in Figure 2.1 (c) and (d) largely indicates possible nonstationarity. In contrast, tight bands of points exhibited in Figure 2.1 (a) and (b) or (e) and (f) are a sign of stationarity. The euclidean distances $u$ and $h$ in Figure 2.1(a-f) are calculated using the respective dimensions from each of the three settings. It is seen that the goodness-of-fit of the MLE of the covariance model in the expanded dimensional space is much improved over the fit in the reduced dimensional space.

### 2.3.2 Numerical results

Table 2.1 reports the MSPE and log score corresponding to each parameter setting for the separable model and Table 2.2 for the nonseparable model. In the separable case, we see that across all parameter combinations, MSPE and log score are the smallest for the true model, but the nonstationary model we propose is the most comparable. The BSZ+ST model is the third best while the stationary and BSZ models perform the worst. Improvements in MSPE and log score of the nonstationary model over others are similar across all weak, moderate and strong spatial and temporal correlations. It is interesting to note that the BSZ model performs worse than the stationary model when there is temporal correlation present, and the discrepancy is more obvious when the temporal correlation is stronger. This indicates that the BSZ model can be insufficient for a spatio-temporal data set.

For the nonseparable case, we again see that the nonstationary model is most comparable to the true model in terms of both MSPE and log score. We also see that the nonstationary model consistently outperforms the BSZ+ST across all parameter

Table 2.1: Comparison between nonstationary models with other models using separable space-time covariance function.

| Model | Dimensions | $\phi_t = \phi_s = 0.3$ | | $\phi_t = \phi_s = 0.5$ | | $\phi_t = \phi_s = 0.8$ | |
|---|---|---|---|---|---|---|---|
| | | MSPE (SE) | LS (SE) | MSPE (SE) | LS (SE) | MSPE (SE) | LS (SE) |
| True | $([t^1, t^2], [s^1, s^2, s^3])$ | 0.043 | -0.676 | 0.106 | 0.261 | 0.222 | 1.046 |
| | | (0.0006) | (0.0097) | (0.0006) | (0.0097) | (0.0028) | (0.01) |
| NonStat | $([t^1, \boldsymbol{w}], [s^1, s^2, \boldsymbol{z}])$ | 0.065 | -0.072 | 0.106 | 0.771 | 0.274 | 1.364 |
| | | (0.0013 ) | (0.0178) | (0.0013) | (0.0178) | (0.0042) | (0.0149) |
| BSZ+ST | $(t^1, [s^1, s^2, \boldsymbol{z}])$ | 0.073 | 0.176 | 0.167 | 1.018 | 0.313 | 1.629 |
| | | (0.0014) | (0.0168) | (0.0029) | (0.0167) | (0.0049) | (0.0151) |
| BSZ | $([s^1, s^2, \boldsymbol{z}])$ | 0.271 | 1.472 | 0.408 | 1.921 | 0.555 | 2.206 |
| | | (0.0055) | (0.0188) | (0.0067) | (0.0193) | (0.009) | (0.0181) |
| Stationary | $(t^1, [s^1, s^2])$ | 0.111 | 0.594 | 0.235 | 1.352 | 0.409 | 1.913 |
| | | (0.0019) | (0.0174) | (0.0019) | (0.0174) | (0.006) | (0.0145) |

*LS represents log score. In the column of Dimensions, $\boldsymbol{z}$ is the estimated dimension for $s^3$ and $\boldsymbol{w}$ is the estimated dimension for $t^2$.

Table 2.2: Comparison between nonstationary models with other models using nonseparable space-time covariance function.

| $\beta$ | Model | $a = 2, c = 2$ | | $a = 2, c = 1$ | | $a = 7, c = 2$ | | $a = 7, c = 1$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | MSPE (SE) | LS (SE) | MSPE (SE) | LS (SE) | MSPE (SE) | LS (SE) | MSPE (SE) | LS (SE) |
| | True | 0.520 | 2.037 | 0.393 | 1.729 | 0.722 | 2.435 | 0.534 | 2.102 |
| | | (0.006) | (0.012) | (0.005) | (0.013) | (0.009) | (0.011) | (0.007) | (0.012) |
| 0.5 | Nonstat | 0.570 | 2.204 | 0.493 | 2.045 | 0.772 | 2.552 | 0.626 | 2.317 |
| | | (0.008) | (0.015) | (0.008) | (0.018) | (0.010) | (0.013) | (0.009) | (0.015) |
| | BSZ+ST | 0.601 | 2.285 | 0.512 | 2.098 | 0.794 | 2.559 | 0.631 | 2.291 |
| | | (0.008) | (0.013) | (0.008) | (0.017) | (0.010) | (0.013) | (0.009) | (0.014) |
| | Stationary | 0.650 | 2.396 | 0.604 | 2.313 | 0.853 | 2.670 | 0.811 | 2.612 |
| | | (0.008) | (0.013) | (0.008) | (0.014) | (0.010) | (0.012) | (0.008) | (0.011) |
| | True | 0.533 | 2.069 | 0.418 | 1.800 | 0.730 | 2.450 | 0.547 | 2.133 |
| | | (0.006) | (0.012) | (0.005) | (0.013) | (0.009) | (0.011) | (0.007) | (0.012) |
| 1 | Nonstat | 0.587 | 2.242 | 0.519 | 2.107 | 0.775 | 2.552 | 0.640 | 2.348 |
| | | (0.008) | (0.014) | (0.009) | (0.018) | (0.009) | (0.012) | (0.004) | (0.014) |
| | BSZ+ST | 0.612 | 2.305 | 0.532 | 2.138 | 0.803 | 2.568 | 0.641 | 2.311 |
| | | (0.008) | (0.013) | (0.008) | (0.016) | (0.010) | (0.012) | (0.009) | (0.014) |
| | Stationary | 0.656 | 2.407 | 0.620 | 2.343 | 0.857 | 2.675 | 0.790 | 2.583 |
| | | (0.008) | (0.013) | (0.009) | (0.015) | (0.010) | (0.012) | (0.010) | (0.013) |

*The dimensions of each model and the notation of LS follow those in Table 2.1.

Table 2.3: Comparison between nonstationary models estimated with different amounts of data, stationary models, and true models. The nonstationary data are generated based on the separable model (2) with parameters $\sigma^2 = 1$, $\phi_t = \phi_s = 0.5$ and the expanded dimensions described in Section 2.3.

| Model | MSPE (SE) | Log Score (SE) |
|---|---|---|
| Stationary | 0.457 | 2.168 |
| | (0.007) | (0.020) |
| $\text{Nonstationary}_{900}$ | 0.402 | 1.840 |
| | (0.010) | (0.026) |
| $\text{Nonstationary}_{2,500}$ | 0.344 | 1.649 |
| | (0.004) | (0.015) |
| $\text{Nonstationary}_{all}$ | 0.253 | 1.311 |
| | (0.004) | (0.019) |
| True | 0.235 | 1.109 |
| | (0.003) | (0.013) |

$\text{Nonstationary}_n$ is the nonstationary model fitted with $n$ selected observations.

combinations except when there is low temporal correlation and strong spatial correlation as with $a = 7$ and $c = 1$. However, the improvement is less prominent compared to the separable case, in particular when comparing the log scores. In the case of $a = 7,\ c = 1$, the two models perform very similarly. This result is not surprising as when the temporal correlation is relative weaker than the spatial correlation, modeling of the temporal correlation, whether stationary or nonstationary, becomes less important. Therefore, the BSZ+ST model is nearly as sufficient as the nonstationary model in this case. The stationary model again consistently performs the worst across all parameters combinations. The pattern of the results seems insensitive to the value of the separability parameter.

To explore the efficiency of the sampling method for larger datasets described in Section 2.2.4, 100 spatio-temporal processes are simulated at $n = 100$ spatial

locations and $T = 100$ time points using the separable covariance function (2.2.2) with $\sigma^2 = 1$ and $\phi_t = \phi_s = 0.5$. For each simulated dataset, we sampled 900 of the observations by randomly selecting 30 spatial locations from the 100 locations, and 30 time points from the 100 time points. We also sampled 2,500 observations by sampling 50 spatial locations and 50 time points. Details of the random selection of locations and time points are described in Section 2.2.4. Then using the 900, 2,500 and all of the observations, we estimate the latent dimensions $\boldsymbol{w}(t)$ and $\boldsymbol{z}(\boldsymbol{s})$, fit the nonstationary model (2.2.2) and compute the MSPE and log score of 100 randomly dropped observations from the whole dataset. The results are reported in Table 2.3. We see a gradual improvement in prediction when more and more observations are used to estimate the latent dimensions. Even with only 900 of the observations, the nonstationary model has significantly reduced the MSPE of the stationary model by 14%. With 2,500 of the observations, the nonstationary model reduces the MSPE of the stationary model by 25%. Figure 2.2 compares the true latent dimensions $s_{3i}$ and $t_{2j},\ i = 1, \ldots n,\ j = 1, \ldots T$ with the estimated dimensions $\boldsymbol{w}$ and $\boldsymbol{z}$ using only 900 observations and those found using all data points. It is seen that the estimates using 900 observations are already similar to those using the whole data set and both share the same pattern as the true latent dimensions. These results show that strategically taking a subset of the data could be used to efficiently apply dimension expansion to a large space-time dataset.

## 2.4 Application

### 2.4.1 Illinois Wind Speed Data

Daily maximum wind speed from 2010 was obtained from Illinois State Water

Table 2.4: Monthly MSPE and log score of using different covariance models

| Month | Stationary | | Nonstationary $(p, q^*)$ | | BSZ | | BSZ+ST | |
|---|---|---|---|---|---|---|---|---|
| | MSPE | LS | MSPE | LS | MSPE | LS | MSPE | LS |
| 1 | 0.73 | 3.80 | 0.51(2,1) | 2.04 | 0.56 | 2.11 | 0.51 | 2.03 |
| 2 | 0.90 | 4.67 | 0.51(1,1) | 1.96 | 0.53 | 2.04 | 0.50 | 2.00 |
| 3 | 0.38 | 2.61 | 0.32(1,1) | 1.50 | 0.32 | 1.52 | 0.32 | 1.51 |
| 4 | 0.30 | 2.60 | 0.24(1,1) | 1.25 | 0.26 | 1.19 | 0.25 | 1.34 |
| 5 | 0.47 | 3.20 | 0.41(3,1) | 1.76 | 0.43 | 1.80 | 0.42 | 1.79 |
| 6 | 0.74 | 3.78 | 0.57(1,1) | 2.16 | 0.57 | 2.17 | 0.57 | 2.17 |
| 7 | 1.04 | 4.68 | 0.72(2,1) | 2.36 | 0.74 | 2.37 | 0.74 | 2.37 |
| 8 | 1.29 | 4.92 | 0.91(2,1) | 2.71 | 0.90 | 2.69 | 0.92 | 2.71 |
| 9 | 0.67 | 3.87 | 0.51(1,1) | 1.96 | 0.50 | 2.00 | 0.51 | 1.98 |
| 10 | 0.61 | 3.86 | 0.35(1,1) | 1.67 | 0.38 | 1.78 | 0.35 | 1.67 |
| 11 | 0.45 | 3.13 | 0.31(2,1) | 1.46 | 0.31 | 1.48 | 0.31 | 1.45 |
| 12 | 0.99 | 4.40 | 0.67(1,1) | 2.33 | 0.70 | 2.43 | 0.67 | 2.33 |

*$p$ and $q$ are the number of estimated latent dimensions of $\boldsymbol{z}$ and $\boldsymbol{w}$.

Survey (*http://www.isws.illinois.edu*). Wind speed is known to typically have nonstationary properties across a large area such as the state of Illinois (Nott and Dunsmoir, 2002). The wind measurements were recorded in miles per hour from 19 weather stations across Illinois and a map of the stations is shown in Figure 1.1. Due to large variation among stations, the data was scaled by location and the overall mean is removed before analysis. Since dimension expansion can become computationally inefficient with $T$ as large as 365, We fit four models (Nonstationary, Stationary, BSZ, BSZ+ST) as described in Section 2.3 using data from each month in 2010. We choose a separable covariance model which is a product of an AR(1) temporal correlation function and an exponential spatial covariance function,

$$C(h, u) = \sigma^2 \rho^{|u|} \exp(\phi h),$$

where $\rho \in (0, 1)$ measures how quickly the temporal correlation decays. The spline

penalty and lasso penalty parameters were optimized for each month separately. It is seen that the choice of those parameters and even the number of learned dimensions vary month by month.

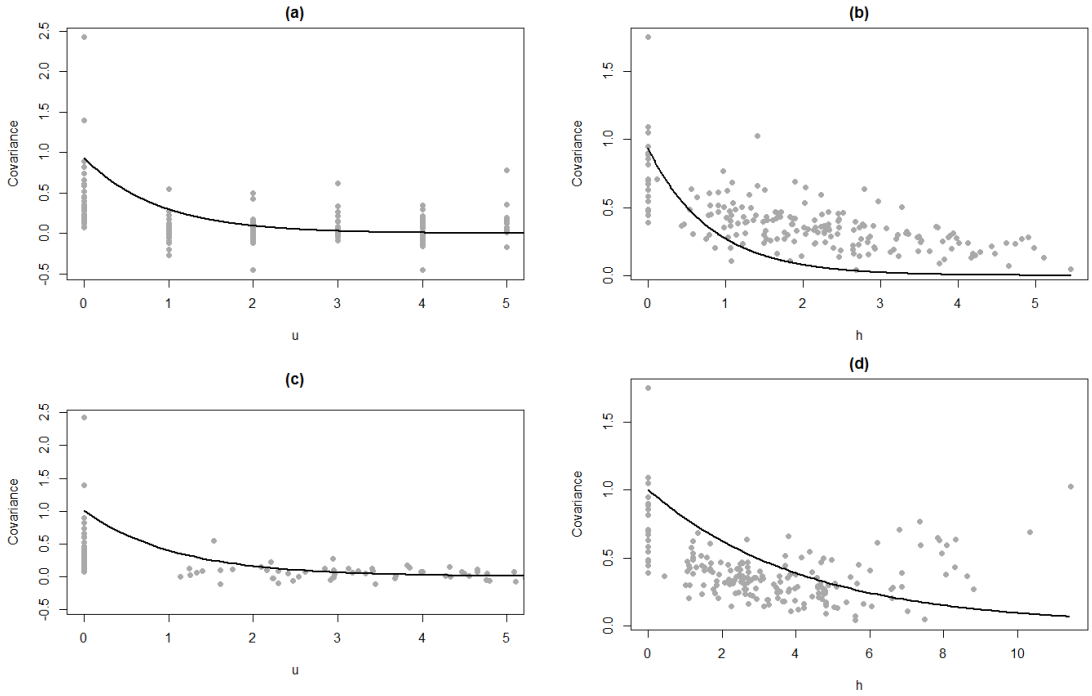


Figure 2.3: (a)-(b)Empirical temporal and spatial covariances with observed dimensions and (c)-(d) estimated dimensions for May.

Table 2.4 shows MSPE and log scores for drop-one predictions for each month, based on the four models. For all months, using learned dimensions in both time and space significantly improves the MSPE and log score, compared to using the observed dimensions. Among the three models using the learned dimensions (Nonstationary, BSZ, BSZ+ST), our method either outperforms or is comparable to the other two. It is worth noting that months that show all methods are comparable such as June, August and September, tend to have either very weak to no temporal correlation or both relatively low temporal and spatial correlation within that month. Figure 2.3 shows examples of the MLE covariance model fits for the month of May which

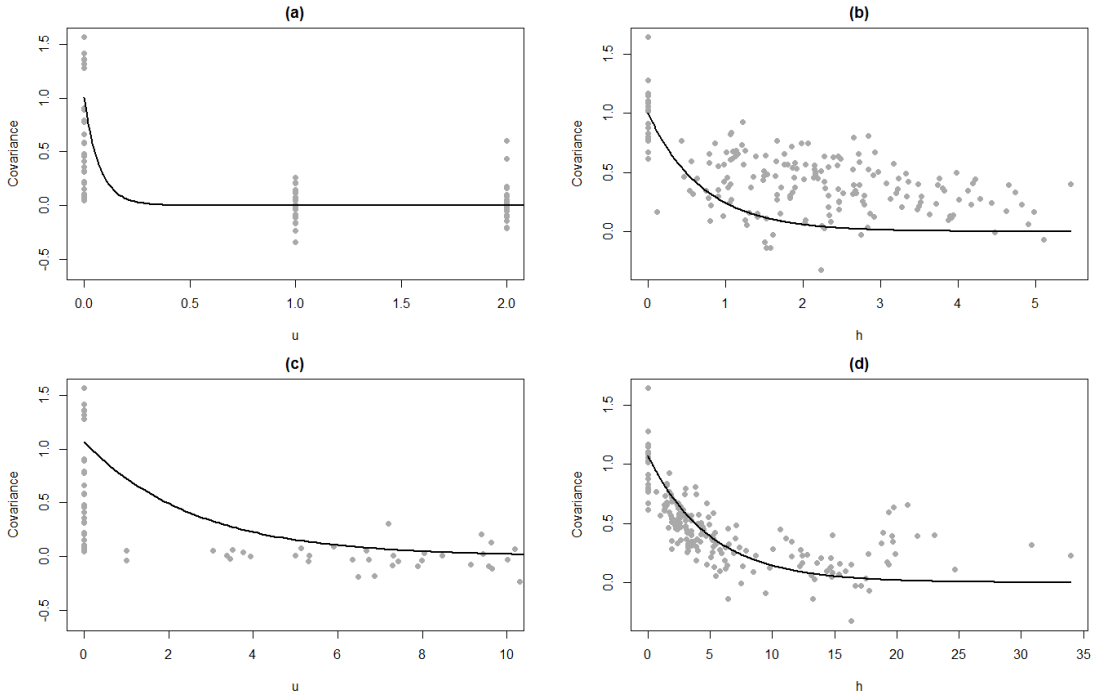


Figure 2.4: (a)-(b)Empirical temporal and spatial covariances with observed dimensions and (c)-(d) estimated dimensions for September.

has relatively moderate weak temporal correlation and Figure 2.4 with the additional dimensions for September, which has very weak temporal correlation. We see a larger improvement in the purely spatial covariance fits, which we believe is due to the overall presence of stronger correlation in space than in time. Therefore, for months that carry very weak temporal correlation, the improvement of using nonstationary model over the space-time that is nonstationary in space but stationary in time is only marginal.

### 2.4.2 Streamflow data

To further illustrate that our nonstationary space-time covariance model can be useful in practice, we apply it to a streamflow dataset. Daily average discharge of streams from 28 stations in Northern Illinois and Wisconsin during June through

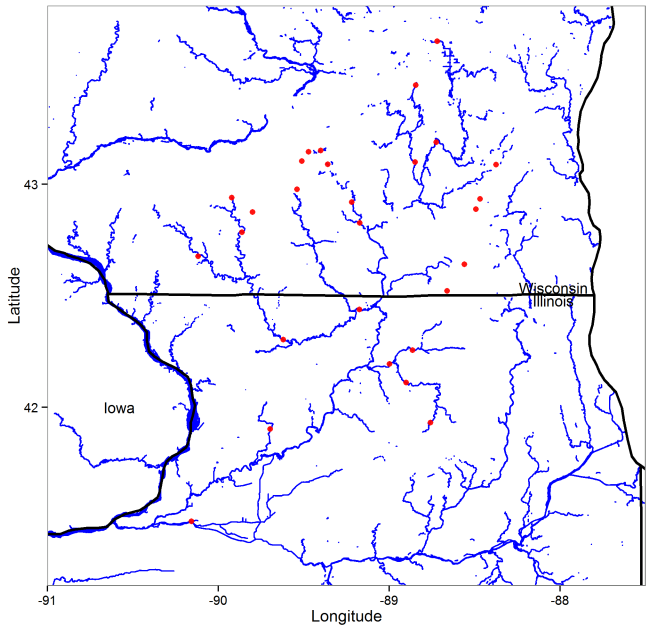


Figure 2.5: Station Locations on streams in Illinois and Wisconsin.

October 2013 was obtained from the US Geological Survey, see Figure 2.5 for a station map. The discharge, measured in $ft^3/s$, is the volume of water passing through any given point in the stream. Due to lake and urbanization effects near Madison, WI and Lake Koshkonong on Rock River, Rock River stations below Lake Koshkonong and with extreme drainage areas were excluded from consideration. A stream is a form of catchment or drainage basin, i.e. an extent of land where surface water from rain, snow melt, etc. converges and drains to a lower elevation. Thus, the discharge at a station depends on the water level at an unknown point of higher elevation upstream within the drainage basin of that station. Considering upstream elevation as a latent dimension and only observing latitude and longitude coordinates of stations, we apply dimension expansion in an attempt to mimic the effect of elevation upstream of each station.

We first take a log transformation of the data and remove the station mean of

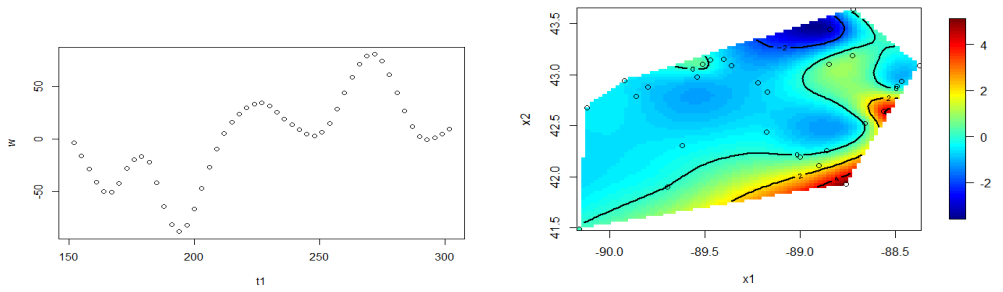


Figure 2.6: Estimated latent temporal dimensions $\boldsymbol{w}$ (upper) and spatial dimensions $\boldsymbol{z}$ (lower).

2013. We then apply dimension expansion with specified separable exponential model (2.2.2). The nonstationary model expands time and space dimensions to $[t^1, w] \in \mathbb{R}^2$ and $[s^1, s^2, z] \in R^3, i = 1, \ldots n,\ j = 1, \ldots T$, while the stationary model uses only the observed dimensions $t^1 \in \mathbb{N}$ and $[s^1, s^2] \in \mathbb{R}^2$. To learn the additional dimensions, every 3rd day within this time frame was used to reduce the number of time points from 151 to 51 and to reduce the sample size from 4,228 to 1,428. Additional dimensions for time and space, $\boldsymbol{w}(t)$ and $\boldsymbol{z}(\boldsymbol{s})$ respectively, were estimated with optimal tuning parameters: $\lambda_{t,1} = 4, \lambda_{s,1} = 0.08, \lambda_{t,2} = 10e^{-5}$, and $\lambda_{2,s} = 10e^{-4}$, which were chosen by minimizing the drop-one hundred MSPE.

Figure 2.6 shows the estimated latent dimensions, $\boldsymbol{w}(t)$ and $\boldsymbol{z}(\boldsymbol{s})$. The relatively large magnitudes of $\boldsymbol{w}$ and $\boldsymbol{z}$ indicate that their influence is not negligible, meaning we would expect a great prediction improvement by our model over the stationary model. We then fit the model using the estimated latent dimensions and compare the MSPE and log score of drop-one hundred predictions of our nonstationary model with BSZ+ST, BSZ and stationary models. To make all models comparable, the latter three models follow the basic form of (2.2.2). Our nonstationary model performs best with MSPE of 0.131 and log score of 0.216 while the stationary model performs the worst with MSPE of 0.165 and log score of 1.15. The BSZ+ST model is the second

best with MSPE of 0.149 and log score of 0.471 and BSZ performs similarly with MSPE 0.140 and log score of 0.514. Although the MSPEs of all models except for the stationary case seem similar, the log score is notably smaller for the nonstationary model. This indicates that modeling nonstationarity in both space and time can improve the prediction variance estimation for certain spatio-temporal data.

The density plot of the streamflow data shows a weak sign of outliers so we additionally carry out the robust estimation by employing the robust variogram estimator proposed by Cressie and Hawkins (1980). The results are very similar to the above suggesting robust estimation may not be necessary for this dataset. We also applied the nonstationary space-time models in Ma (2002) and Garg et al. (2012) to this data set. The drop-100 MSPE from those two methods are 0.264 and 0.495 and log scores 1.572 and 543.8 , respectively. We conjecture that their poor performance could be because the model in Ma (2002) is too parsimonious for this data as it only involves a few parameters, while the model in Garg et al. (2012) may have overfitting issues as it involves many parameters but no penalty on the flexibility of the model is discussed in their model fitting procedure.

## 2.5 Discussion

Many spatio-temporal datasets in the natural sciences exhibit nonstationary dependence structures in both space and time due to various reasons, although most existing space-time covariance models assume second order stationarity. Dimension expansion in both space and time offers an easy way to allow for nonstationarity while still enjoying the abundance of traditional stationary covariance models. We investigate the estimation of parameters and latent dimensions in space and time for separable and nonseparable space-time covariance models and illustrate our method

in simulation studies and a real data application. It is seen that our method can significantly improve prediction compared to using observed dimensions when the data are nonstationary. Therefore, the comparison may give rise to a crude way to test for stationarity, an alternative to the test proposed by Jun and Genton (2012).

There is no definitive answer to the question of how many latent dimensions are sufficient to attain stationarity. In practice, the number of expanded dimensions $p$ and $q$ should be chosen as the fewest possible to construct a stationary space. This can be explored by a trial and error strategy, gradually increasing $p$ and $q$. The lasso penalty in equations (2.2.1) and (2.2.5) will impose the coordinates of unnecessary additional dimensions to be close to zero, indicating the best choice. In our experience, we have found one to two extra dimensions to be sufficient.

Our paper mainly focuses on modeling the covariance structure at fine resolution and we illustrate the method using small datasets. Dimension expansion can be computationally expensive as the number of extra dimensions, $p$ or $q$, or the number of space-time locations, $n$ or $T$ increases. For large spatial data, Bornn et al. (2012) suggested the gradient projection method of Kim et al. (2006) for the optimization. Without resorting to more complex optimization algorithms, we propose a strategic approach to reduce the computation by estimating the latent dimensions using only a subset of observations. This method shows great potential in practice. Another promising approach to deal with computation for large data is the binned method proposed by Kang et al. (2010), which divides the entire data set into bins, treating each bin as an observation and the locations within each bin as replicates.

The identifiability of latent dimensions is not rigorously investigated here. Bornn et al. (2012) briefly discussed the conditions under which the precise latent dimensions are identifiable, but they emphasized that dimension expansion mainly focuses on approximating a real process rather than the estimation of the latent dimensions.

We agree that the main merit of using latent dimensions lies in their practical advantages and that the estimation method is imperfect. Furthermore, we note that the estimation of latent dimensions depends on the choice of the covariance model. Given a different choice, the estimates of latent dimensions will likely be different. Hence the method of dimension expansion is essentially an attempt of approximating the nonstationarity based on a given covariance model.

# Chapter 3

# Spatially Varying Autoregressive Models for Prediction of New HIV Diagnoses

## 3.1 Background

Human immunodeficiency virus (HIV) infections are life-changing events to those inflicted and if left untreated can lead to acquired immunodeficiency disease (AIDS). The Centers for Disease Control and Prevention (CDC) has reported that although nationally the number of newly diagnosed cases of HIV has declined by 19% in the last decade, progress has been uneven. For example, some demographic groups, such as African Americans, and some geographic regions, such as the south of the US, have shown slower declines, if any, exhibiting consistently high HIV rates [1]. Yet other regions that were not traditionally affected have seen dramatic outbreaks, such as Scott county, Indiana. Regional prediction of disease is central to orchestrating appropriate

[1] *http://www.cdc.gov/hiv*

public health responses. The National HIV/AIDS Strategy [2] identifies a key goal of intensifying efforts in the communities with the greatest concentration of HIV cases. Developing models to predict future diagnoses should allow health departments to intervene before the surge in new diagnoses occurs. Although admittedly imperfect, earlier intervention offers the possibility of reaching people living with HIV sooner, and of improving health and decreasing infectiousness through timely treatment. Failures to anticipate the sudden increases in HIV in Scott County, Indiana, for example, might also be averted with regional predictions.

In this article, we will focus on the prediction of new HIV diagnosis rates at the county level using publicly available data. It is well-known that the presence of infection in a region is partly influenced by the social and economic demographics of its population. Known demographic variables linked to health disparities include but are not limited to education, income, health-care access, sexual orientation, and ethnicity. However, demographics alone are insufficient to explain the entire variability in the HIV data. After the effects due to the demographic covariates are removed, the spread of infection across the U.S. still exhibits strong spatially and temporally varying patterns as values among neighboring regions and time periods tend to be similar. It is therefore necessary to have appropriate modeling techniques that incorporate both the variability due to space-time dependencies and the variability due to demographic covariates. Another major challenge that comes with the prediction of HIV lies in the rarity of the disease, leading to few to no incidents in many regions. This sparsity strongly encourages prediction methods to take advantage of the spatial and temporal dependency structures so that the statistical inference at one location can borrow strength from neighboring regions in both space and time. All aforementioned challenges motivate us to investigate efficient statistical models with the goal

[2] *https://www.aids.gov/federal-resources/national-hiv-aids-strategy/overview/*

of predicting new HIV diagnoses.

Incidences of disease are most popularly modeled as Poisson or Binomial random variables with mean functions dependent on relative risk, expressed as the ratio of observed risk to expected risk. The expected risk is calculated by standardizing the observed number of people at risk in each combination of region and period by age, gender or other categorization. This standardization can be performed either externally or internally, depending on whether the information retrieved for standardization is from another data source or from within the given data. The relative risk is then modeled through a link function, say, log or logit link function. Modeling of the transformed relative risk at region $i$ and period $t$, $\eta_{i,t}$, is the main goal of disease modeling. This goal is usually achieved through an additive linear function which is essentially the sum of linear effects from certain covariates and spatial/temporal/spatio-temporal random effects. The spatial or temporal random effects may be viewed as surrogates for unobserved regional or time-changing covariates.

Spatial dependence for aggregated disease incidence is most commonly modeled through the conditional autoregressive (CAR) model. The intrinsic CAR (ICAR) model proposed by Besag et al. (2011) has been widely used but has a highly restrictive on the correlation structure that is only appropriate for data with strong spatial correlation. To make the model more flexible, Besag et al. (2011) proposed to combine the ICAR model with an additional set of independent random effects to allow for a range of correlation strengths. Cressie (1993) and Stern and Cressie (2000) proposed to have a single set of random effects that still enables varying strengths of spatial correlation to be captured. Leroux et al. (1999) modified Cressie (1993) to make it more theoretically appealing by providing the explicit joint distribution of the spatial observations. Later this model was used in MacNab (2003) to analyze intraventricular hemorrhage incidence rates. Lee (2011) showed that among the above mentioned

models, Leroux et al. (1999) is preferable because it consistently produces the best model fit across a large range of possible spatial correlation strengths.

Spatio-temporal disease models vary widely in the methods used to incorporate the temporal variability and its interactions with spatial variability. Bernardinelli et al. (1995b) and Sun et al. (2000) captured the spatially varying temporal trend using a linear function of time with region-specific intercepts and slopes. Assunção et al. (2001) incorporated an additional quadratic term in time to allow for curved trends. Alternatively, MacNab and Dean (2001, 2002) used B-splines to model the temporal trend independently for each region such that the models share information in time but not across space. Waller et al. (1997) and Xia and Carlin (1998) on the contrary built independent spatial models for each time period. Although such a model is highly flexible, it can quickly become high dimensional by treating each time point separately. To make the temporal evolution less restrictive and to allow observations to share information in both space and time, Knorr-Held (2000) proposed a generalized linear mixed model framework with spatial and temporal main effects, estimated as an ICAR and random walk model respectively, as well as a possible space-time interaction term, expressed as the kronecker product of the main effects. Lagazio et al. (2003) and Schmid and Held (2004) extended this model by incorporating temporally varying covariates, creating age-period-cohort models. Later, Nobre et al. (2005) proposed a dynamic generalized linear mixed model by allowing both the fixed and random effects to evolve over time. Martínez-Beneito et al. (2008) proposed a multivariate autoregressive model that uses a single, spatially-invariant parameter to estimate the disease's changing rate over time and used Besag et al. (2011) to capture the spatial correlation in the innovations of the autoregressive model.

More recently, Cai et al. (2014) built a semiparametric model with spatio-temporally varying regression coefficients which are further decomposed into fixed, spatially vary-

ing and temporally varying components. While the temporally varying components are modeled through a dynamic model, the spatially varying components are modeled via a nonparametric Dirichlet process. Motivated by Knorr-Held (2000), Bauer et al. (2016) decomposed disease risk into purely spatial and purely temporal components and a space-time interaction term. The space-time interaction term is modeled via tensor product splines instead of as a product of the main effects, thus reducing the complexity of the model. Ugarte et al. (2012) applied a similar additive structure to prostate cancer mortality data and modeled the random effects using P-splines. Ruiz-Medina et al. (2014) proposed another nonparametric approach to disease modeling by treating the risk of breast cancer mortality as functional data. Finally, Mercer et al. (2015) incorporated multiple survey effects into generalized linear mixed models to accommodate the particular features of the widely available demographic and health survey data related to modeling child mortality.

While the majority of the aforementioned statistical spatio-temporal models mainly focus on capturing the underlying pattern of disease risk and/or the association between the disease and environmental variables, our goal here is to make predictions for disease rates. We propose a new class of spatially varying autoregressive (SVAR) models for this purpose, and compare them to six models in the popular additive modeling framework summarized in Knorr-Held (2000) and additionally two spatially invariant autoregressive models. Our SVAR models allow the temporal correlation to be location specific while still remaining parsimonious so that the models are governed by only a few parameters. This is particularly important for the HIV data presented here, which only have at most six observations in time for each county, making the individual estimation of an autoregressive model for each county unaffordable. Our models are more interpretable and intuitive for space-time disease data compared to the additive mixed effects models. We will evaluate the prediction capabilities of our

models versus the additive models and spatially invariant models by applying them to the HIV data over three regions, highlighting their strengths and weaknesses.

This chapter is organized as follows: Section 3.2 describes the data, Section 3.3 introduces our SVAR models and presents the competing models, and Section 3.4 presents the prediction results. Finally, Section 3.5 provides a brief conclusion and discusses the implications of the approaches explored. Additional plots and all derivations are deferred to the Appendix.

## 3.2 Data

Annual new HIV diagnosis data from 2008 to 2014 at the county level across the United States is available at *AIDSVu.org*. HIV rates are reported as the number of cases per 100,000 people for a given county. In order to protect the privacy of individuals, HIV rates in a county may be suppressed in any of the following situations: (1) a county has very few cases ($< 5$) or has a small population size ($< 100$), (2) the state health department requested not to release its data to AIDSvu due to re-release agreements with the CDC, and (3) there are no counties in the state such as in Alaska, District of Columbia and Puerto Rico. Due to the rareness of the disease as well as the confidentiality constraints, only 25% of all possible county-time observations across the United States have new diagnoses available in the given time frame. Figure 3.1 shows the sparsity of new diagnosis rates across the US in 2012. In this map, negative rates indicate missing values. Hot spots seen on this map agree with reports made by the CDC stating that the highest rates of diagnoses are in the South, the West and the Northeast. For this reason, we focus on three concentrated areas of the US: Florida with 67 counties and 75% of county-time observations available, California with 58 counties and 59% of county-time observations available, and the group of seven New

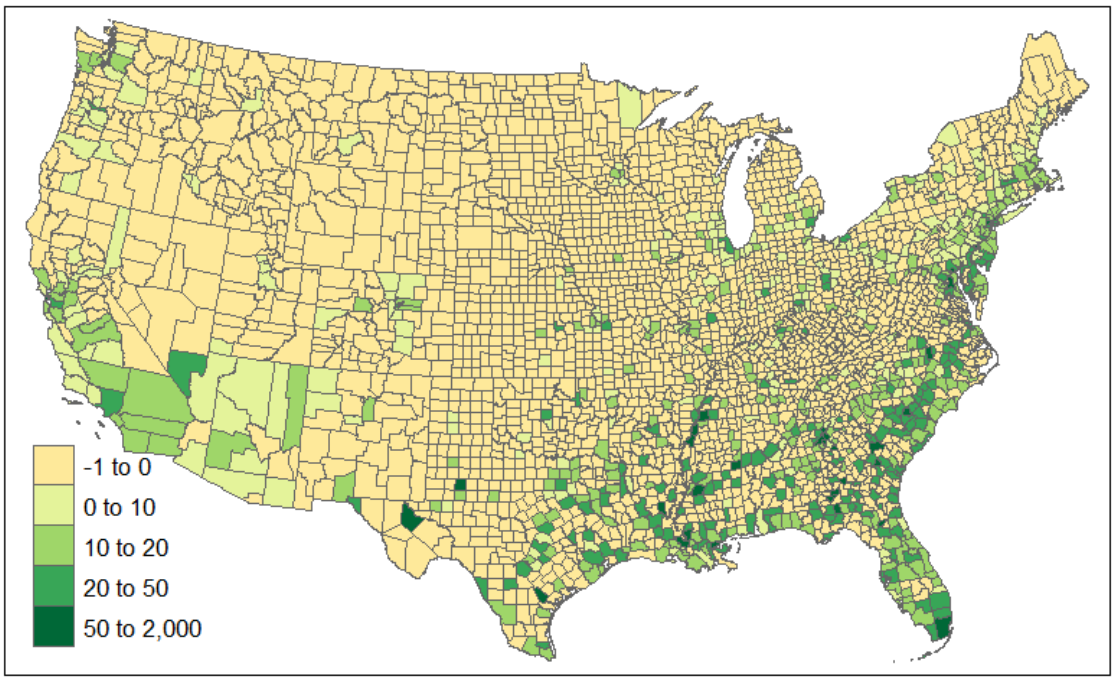


Figure 3.1: 2012 new HIV diagnosis rates in cases per 100,000 across the United States.

England states: Connecticut, Delaware, Maryland, Massachusetts, New Jersey, New York, and Pennsylvania with 199 counties and 74% county-time observations available collectively.

General demographic annual summaries by county such as age, gender, sex at birth, race, house-hold information, and population for the years 2007-2014 were obtained from the United States Census Bureau. Other social and economic variables such as education, marital status, income, and same-sex couple households are also available from the American Community Survey run by the U.S. Census Bureau. We consider demographic variables from the previous year as covariates. Since annual demographic data for all variables is only available for highly populated counties, we use 3-year or 5-year averaged data for those counties with missing annual or 3-year averaged variables respectively. Prevalence of other sexually transmitted infections

such as chlamydia and syphilis are also available from *healthindicators.gov*. Although there are over 100 possible covariates from all available data collection sources, past research suggests that only a small fraction of social and economic variables are significantly related to the spread of HIV. However, as not all variables currently available have been previously considered, we performed our own preliminary analysis for initial variable selection. Using a step-wise regression procedure performed on all US data available, we selected 28 significant demographic variables. The more significant covariates include race, education level, activity level, and number of same sex households, which are known to be related to the epidemic nature of the HIV disease. The effects of these variables on HIV have also been shown in previous HIV studies (e.g., Balaji et al., 2013 and Albarracín et al., 2010).

## 3.3 Model development and comparison

Our goal is to make county level one year ahead predictions of new HIV diagnosis rates based on the observed county level 2008-2014 new HIV diagnoses. For such a short time series, a linear or quadratic temporal trend model may impose too strong an assumption. We therefore choose an autoregressive model with order 1 (AR(1)) structure to model the temporal correlation of each county. However, an obvious challenge with this data set is that it is statistically unreliable to fit individual AR(1) models for each county due to so few observed time points. Moreover, the evolution rates of neighboring counties are likely similar so independent AR(1) models may fail to account for the anticipated similarity of the autoregressive coefficient, $\rho$, in nearby or adjacent counties. The similarity of evolution rates in neighboring counties is verified in an exploratory analysis, where we fit an AR(1) model for each county independently and examine the spatial correlation of the estimated $\rho$s. Details can

be found in the Appendix Section A.1.3.

To accommodate the characteristics of the data and also ensure the model inference is feasible, we propose to jointly model the individual evolution of the rates in each county by modeling the county specific $\rho$ as a spatially dependent random process using the CAR model developed by Leroux et al. (1999). This SVAR modeling strategy has several advantages: first, it allows for flexible county-specific autoregressive coefficients; second, it makes the estimation of an AR(1) model for each county possible by borrowing strength from neighbors since the spatial locations will essentially act as "replicates" in the estimation; and third, it dramatically reduces the rank of the model, easing the potential for overfitting.

Let $Y_{i,t}$ denote the new HIV diagnosis rate for county $i = 1, ..., n$ at year $t = 1, ..., T$. In general, disease modeling first approximates the distribution of raw counts using a Poisson or binomial process with the mean defined as a function of the standardized relative risk, and then focuses on modeling the relative risk strategically using linear mixed models. These models can be put in the framework of generalized linear mixed models (Breslow and Clayton, 1993). However, if the counts are sufficiently large then some type of transformation can be used to attain the normality of the transformed data (Waller et al., 1997). Our data reports $Y_{i,t}$ as the new HIV diagnosis rates in cases per 100,000 rather than raw counts for each county. Moreover, due to the sensitivity of the data, rates below five are suppressed, ensuring that the majority of data are sufficiently large and normality may be achieved by a transformation of $Y_{i,t}$. A transformation can be advantageous over Poisson or binomial approximation because the Gaussian model for the transformed data will allow for direct modeling of overdispersion in the data. For our data, we found that a log transformation works well in transforming data to approximate Gaussian distribution (see Figures A.1-A.4 in the Appendix for the normality plots of the transformed data).

We therefore first take a log transformation of $Y_{i,t}$ and then remove its global mean to obtain a zero mean Gaussian random process $Z_{i,t}$.

We model $Z_{i,t}$ in the first level of the BHM models as below:

$$Z_{i,t} = \eta_{i,t}(\boldsymbol{\beta}, \boldsymbol{\theta}) + \epsilon_{i,t}, \tag{3.3.1}$$

where $\eta_{i,t}$ is a spatio-temporal random process governed by parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

An important feature of the traditional disease modeling framework is that the sampling variance of $Y_{i,t}$ depends on $n_{i,t}$ of county $i$ at time $t$. To achieve this in our modeling framework, we consider the asymptotic properties of the Poisson distribution. As $n_{i,t} \to \infty$,

$$Y'_{i,t} \xrightarrow{D} \mathrm{N}(n_{i,t}r_{i,t}, n_{i,t}r_{i,t}).$$

and since $Y_{i,t} = cY'_{i,t}/n_{i,t},\ c = 100,000$,

$$Y_{i,t} \xrightarrow{D} \mathrm{N}\big(cr_{i,t}, \frac{c^2 r_{i,t}}{n_{i,t}}\big).$$

Transforming $Y_{i,t}$ by taking the natural log, we have

$$\log(Y_{i,t}) - \log(cr_{i,t}) \xrightarrow{D} \mathrm{N}\big(0, \frac{1}{n_{i,t}r_{i,t}}\big)$$

by the delta method. Lastly, we can replace $r_{i,t}$ by its consistent estimator $Y'_{i,t}/n_{i,t} = Y_{i,t}/c$, by Slutsky's theorem, so we have,

$$\log(Y_{i,t}) - \log(cr_{i,t}) \xrightarrow{D} \mathrm{N}\big(0, \frac{c}{n_{i,t}Y_{i,t}}\big).$$

Evidently the asymptotic variance of $\log(Y_{i,t})$ is inversely proportional to $n_{i,t}Y_{i,t}$ and thus the variance of the random error $\epsilon_{i,t}$ should also be inversely proportional to

$n_{i,t}Y_{i,t}$. For this reason, we assume $\epsilon_{i,t} \sim N(0, \sigma^2\mathbf{Q})$ in model (3.3.1), where $\mathbf{Q}$ is an $nT \times nT$ diagonal matrix with the $(it)^{th}$ diagonal entry as $\mathbf{q}_{i,t} = c/(n_{i,t}Y_{i,t})$ for a constant $c = 100,000$.

At the second level, we model $\eta_{i,t}$ using a SVAR model. The basic form of our SVAR model is

$$\eta_{i,t}(\boldsymbol{\beta}, \boldsymbol{\theta}) = X_{i,t-1}^T\boldsymbol{\beta} + \psi_{i,t-1}\rho_i(Z_{i,t-1} - X_{i,t-2}^T\boldsymbol{\beta}), \tag{3.3.2}$$

where $X_{i,t-2}^T\boldsymbol{\beta}$ denotes the linear effects of the previous year's demographic variables at county $i$ on $Z_{i,t-1}$, and $\rho_i \in (-1, 1), i = 1, ...n$, are spatially varying AR(1) coefficients. The coefficient $\psi_{i,t-1} = \frac{\sqrt{q_{i,t}}}{\sqrt{q_{i,t-1}}}$ is added to ensure that $\rho_i$ measures the correlation between two random components, because the variance of $Z_{i,t}$ is proportional to $q_{i,t}$. For all models considered, coefficients in $\boldsymbol{\beta}$ are given independent normal priors.

Depending on the dependency structure in data, we may need to include additional random effects in model (3.3.2). For example, if the spatial correlation in $\rho_i$ is not sufficient to capture the dependency in $Y_{i,t}$, we can add a spatial random effect term $\phi_i$ to the model:

$$\eta_{i,t}(\boldsymbol{\beta}, \boldsymbol{\theta}) = X_{i,t-1}^T\boldsymbol{\beta} + \psi_{i,t-1}\rho_i(Z_{i,t-1} - X_{i,t-2}^T\boldsymbol{\beta}) + \phi_i, \tag{3.3.3}$$

or additionally a spatio-temporal interaction random effect $\delta_{i,t}$:

$$\eta_{i,t}(\boldsymbol{\beta}, \boldsymbol{\theta}) = X_{i,t-1}^T\boldsymbol{\beta} + \psi_{i,t-1}\rho_i(Z_{i,t-1} - X_{i,t-2}^T\boldsymbol{\beta}) + \phi_i + \delta_{i,t}.$$

Note that the major difference between model (3.3.3) and Martínez-Beneito et al. (2008) is that we allow $\rho_i$ to vary in space. Moreover, our spatial random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)^{\mathrm{T}}$ will be governed by the CAR model from Leroux et al. (1999)

rather than Besag et al. (2011) to capture a broader range of spatial correlation strength.

### 3.3.1 Prior specification

The prior specification for $\rho_i$ is challenging because the typical Gaussian CAR prior is not applicable due to the truncated range of $\rho_i$. We therefore propose to model the prior using a copula approach which enables modeling the marginal distribution separately from the dependency structure. By this strategy we are able to model the dependency of $\rho_i$ using a Gaussian random field while still maintaining the flexibility of choosing an approrpriate marginal distribution to respect the truncated nature of $\rho_i$. We first define the joint cumulative distribution function (CDF) of $\rho_1, \cdots, \rho_n$ through a Gaussian copula model:

$$\boldsymbol{C}(\rho_1, \cdots, \rho_n) = C_\Omega(U(\rho_1), \ldots, U(\rho_n)), \tag{3.3.4}$$

where $U(\cdot)$ represents the CDF of the marginal distribution of $\rho_i$, and $C_\Omega(\cdot)$ is a Gaussian copula with covariance matrix $\Omega$ defined by:

$$C_\Omega(u_1, \ldots, u_n) = \boldsymbol{\Phi}_\Omega(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)), \tag{3.3.5}$$

where $\boldsymbol{\Phi}_\Omega(\cdot)$ is the joint CDF of a multivariate Gaussian distribution with zero mean and covariance matrix $\Omega$, and $\Phi^{-1}(\cdot)$ is the inverse CDF of a standard normal random variable.

Here we assume $U(\cdot)$ to be Uniform(-1,1) for simplicity. This choice allows us to

simplify (3.3.4) into

$$\boldsymbol{C}(\rho_1, \cdots, \rho_n) = C_\Omega \left( \frac{\rho_1 + 1}{2}, \ldots, \frac{\rho_n + 1}{2} \right).$$

With (3.3.5), we further have

$$\boldsymbol{C}(\rho_1, \cdots, \rho_n) = \boldsymbol{\Phi}_\Omega \left( \Phi^{-1}(\frac{\rho_1 + 1}{2}), \ldots, \Phi^{-1}(\frac{\rho_n + 1}{2}) \right). \tag{3.3.6}$$

Then the joint prior of $\rho_1, \cdots, \rho_n$ is the density function corresponding to the joint CDF in (3.3.6). Given the density of the Gaussian copula in(3.3.5) is

$$c_\Omega(u_1, \cdots, u_n) = \frac{\boldsymbol{\phi}_\Omega(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n))}{\prod_{i=1}^n \phi(\Phi^{-1}(u_i))},$$

where $\boldsymbol{\phi}_\Omega(\cdot)$ is the density of $\boldsymbol{\Phi}_\Omega(\cdot)$ and $\phi(\cdot)$ is the density of $\Phi(\cdot)$, we obtain the prior for $\rho_1, \cdots, \rho_n$ as:

$$\pi(\rho_1, \cdots, \rho_n) = \frac{\boldsymbol{\phi}_\Omega \left( \Phi^{-1}(\frac{\rho_1+1}{2}), \ldots, \Phi^{-1}(\frac{\rho_n+1}{2}) \right)}{\prod_{i=1}^n \frac{1}{2} \phi \left( \Phi^{-1}(\frac{\rho_i+1}{2}) \right)}. \tag{3.3.7}$$

In practice, we may replace 1 with 0.9999 and 2 with 1.9998 in (3.3.7) to avoid the singularity issue at the boundary of $\rho_i$'s domain. The marginal distribution for $\rho_i$ can be assumed to be any other distribution defined on (-1,1), for example, a truncated normal distribution. However, in such cases a more complex form of the prior $\pi(\cdot)$ will be expected which may significantly increase the computational burden. Given no a priori information for $\rho_i$ the uniform distribution seems a reasonable choice.

Examples of using copula-based methods in a geostatistical setting can be seen in Bárdossy (2006) and Kazianka and Pilz (2010). We choose the covariance matrix $\Omega$ in the Gaussian copula model to follow the structure of the Leroux et al. (1999) CAR

model with variance $\tau_\rho^2$ and spatial correlation parameter $\lambda_\rho$:

$$\Omega = \tau_\rho^2[(1-\lambda_\rho)I + \lambda_\rho R]^{-1},$$

where $R$ denotes a neighborhood matrix with the $i^{th}$ diagonal element as the total number of neighbors for county $i$ while the $(i,j)^{th}$ off-diagonal element is -1 if counties $i$ and $j$ share a border and 0 otherwise. In the Leroux et al. (1999) model, a smaller/larger $\lambda_\rho$ indicates weaker/stronger spatial correlation with $\lambda_\rho = 0$ corresponding to complete spatial independence and $\lambda_\rho = 1$ to the ICAR model. The priors for the parameters of $\rho_i$ are deferred to Section 3.3.2.

Note model (3.3.2) can be extended to include a second order autoregressive term if necessary, although this extension is not appropriate for our data due to the very few time points available. With a spatially varying AR(2) model, we will have both the first and second order coefficients, $\boldsymbol{\rho}_1 = \{\rho_{11}, \cdots, \rho_{1n}\}$ and $\boldsymbol{\rho}_2 = \{\rho_{21}, ..., \rho_{2n}\}$, at $n$ spatial locations. Taking into account the stationarity constraints on each pair of $(\rho_{1k},\ \rho_{2k})$ for $k = 1, \ldots, n$, i.e., $|\rho_{2k}| < 1$ and $|\rho_{1k}| < 1 - \rho_{2k}$, we can model the prior for $(\boldsymbol{\rho}_1,\ \boldsymbol{\rho}_2)$ jointly using Gaussian copulas, $C_{\Omega_1}(\cdot)$ and $C_{\Omega_2}(\cdot)$, which depend on covariance matrices $C_{\Omega_1}(\cdot)$ and $C_{\Omega_2}(\cdot)$ respectively. We assume marginals $\rho_{2k} \sim$ Unif(-1, 1) and $\rho_{1k}|\rho_{2k} \sim$ Uniform($\rho_{2k} - 1$, $1 - \rho_{2k}$) for simplicity. The joint prior can then be specified as

$$\pi(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2) = \pi(\boldsymbol{\rho}_1|\boldsymbol{\rho}_2)\pi(\boldsymbol{\rho}_2),$$

where by (3.3.7),

$$\pi(\boldsymbol{\rho}_2) = \frac{\boldsymbol{\phi}_{\Omega_2}\left(\Phi^{-1}(\frac{\rho_{21}+1}{2}), \ldots, \Phi^{-1}(\frac{\rho_{2n}+1}{2})\right)}{\prod_{i=1}^{n} \frac{1}{2}\phi\left(\Phi^{-1}(\frac{\rho_{2i}+1}{2})\right)},$$

and

$$\pi(\boldsymbol{\rho}_1|\boldsymbol{\rho}_2) \sim \frac{\boldsymbol{\phi}_{\Omega_1}\left(\Phi^{-1}(\frac{\rho_{11}+1-\rho_{21}}{2-2\rho_{21}}),\ldots,\Phi^{-1}(\frac{\rho_{1n}+1-\rho_{2n}}{2-2\rho_{2n}})\right)}{\prod_{i=1}^{n}\frac{1}{2}\phi\left(\Phi^{-1}(\frac{\rho_{1i}+1-\rho_{2i}}{2-2\rho_{2i}})\right)}.$$

### 3.3.2 Other competing models

Most models developed for spatio-temporal disease aim at describing the disease pattern and their dependency structure rather than at prediction. For example, the spatio-temporal models in Knorr-Held (2000) as well as the models reviewed in López-Quílez and Muñoz (2009) are all proposed as generalized linear mixed models for the purpose of disease modeling. These models mainly differ from our models in the construction of $\eta_{i,t}$ specified in (3.3.2). Below lists eight different forms of $\eta_{i,t}$ that will be compared to our SVAR models:

1. $X_{i,t-1}^T\boldsymbol{\beta}$,
2. $X_{i,t-1}^T\boldsymbol{\beta} + \phi_i$,
3. $X_{i,t-1}^T\boldsymbol{\beta} + \alpha_t$,
4. $X_{i,t-1}^T\boldsymbol{\beta} + \delta_{i,t}$,
5. $X_{i,t-1}^T\boldsymbol{\beta} + \alpha_t + \phi_i$,
6. $X_{i,t-1}^T\boldsymbol{\beta} + \alpha_t + \phi_i + \delta_{i,t}$,
7. $X_{i,t-1}^T\boldsymbol{\beta} + \psi_{i,t-1}\rho(Z_{i,t-1} - X_{i,t-2}^T\boldsymbol{\beta})$,
8. $X_{i,t-1}^T\boldsymbol{\beta} + \psi_{i,t-1}\rho(Z_{i,t-1} - X_{i,t-2}^T\boldsymbol{\beta}) + \phi_i$,

where $X_{i,t-1}^T\boldsymbol{\beta}$ is the same as in model (3.3.2), $\phi_i$ denotes the spatial random effect of county $i$, $\alpha_t$ the temporal random effect of year $t$, and $\delta_{i,t}$ the space-time interaction effect. Models 2 through 6 are constructed by adapting Knorr-Held (2000)'s frame-

work to be appropriate for our prediction priority and to make them more comparable to our proposed models. Specifically, we replace its temporal random walk prior with an AR(1) structure and its spatial effects prior following Besag et al. (2011) with that of Leroux et al. (1999). These modifications add more flexibility to the models as the strength of the temporal and spatial correlations in the data can be directly inferred by the correlation parameters in $\alpha_t$ and $\phi_i$, respectively. Models 7 and 8 are specified as invariant autoregressive models across spatial locations, resembling those in Martínez-Beneito et al. (2008).

We specify the priors for $\boldsymbol{\phi} = (\phi_1, ..., \phi_n)^{\mathrm{T}}$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_T)^{\mathrm{T}}$ and $\boldsymbol{\delta} = (\delta_{11}, ...\delta_{nT})^{\mathrm{T}}$ as the following:

$$\boldsymbol{\phi} \sim N(0, \Sigma_\phi),\ \boldsymbol{\alpha} \sim N(0, \Sigma_\alpha),\ \boldsymbol{\delta} \sim N(0, \Sigma_\delta),$$

where $\Sigma_\phi$ follows the structure of the Leroux et al. (1999) CAR model:

$$\Sigma_\phi = \tau_\phi^2[(1 - \lambda_\phi)I + \lambda_\phi R]^{-1},$$

where $\tau_\phi^2$ denotes the effect variance and $\lambda_\phi \in [0, 1]$ the strength of the spatial correlation. The covariance matrix $\Sigma_\alpha = \tau_\alpha^2 \rho_\alpha^D$ is assumed to follow an AR(1) structure with variance parameter $\tau_\alpha^2$, temporal correlation parameter $\rho_\alpha$ and $T \times T$ temporal distance matrix $D$. We specify $\Sigma_\delta$ as the Kronecker product of AR(1) temporal correlation and Leroux et al. (1999) spatial correlation structures:

$$\Sigma_\delta = \tau_\delta^2 \rho_\delta^D \otimes [(1 - \lambda_\delta)I + \lambda_\delta R]^{-1},$$

where $\tau_\delta^2$ is the variance parameter, $\rho_\delta$ and $\lambda_\delta$ denote the temporal and spatial correlation parameters respectively. This indicates a separable space-time covariance

structure. In Models 2 through 6, as well as in Models 8 and (3.3.3), there are clear identifiability issues since the overall trend can be absorbed by either $\alpha_t$, $\phi_i$, or even $\delta_{i,t}$. To remedy this, we follow the simple suggestion made by Knorr-Held (2000) to center $\boldsymbol{\alpha}, \boldsymbol{\phi}$, and $\boldsymbol{\delta}$ to mean zero after each sampling iteration. A small caveat for Models 4 and 6 is that if $\rho_\delta$ and $\lambda_\delta$ are both zero, then $\delta_{i,t}$ is indistinguishable from $\epsilon_{i,t}$. Thus, in implementing Models 4 and 6, estimates of $\rho_\delta$ and $\lambda_\delta$ should be carefully monitored.

### 3.3.3 Hyperpriors

The temporal and spatial dependence in the correlation structures of the random effects are defined through hyperparameters which will be estimated via MCMC sampling. The variance parameters, $\sigma^2$, $\tau_\alpha^2$, $\tau_\phi^2$, $\tau_\delta^2$, and $\tau_\rho^2$ are all given a semi-conjugate Inverse Gamma hyperprior $IG(a, b)$ and can be sampled via the Gibbs sampler. The correlation parameters $\lambda_\phi$, $\lambda_\delta$, and $\lambda_\rho$ are given Uniform(0,1) hyperpriors while $\rho_\alpha$ and $\rho_\delta$ are given Uniform(-1,1) hyperpriors and are each sampled via the Metropolis-Hastings algorithm. Following Bernardinelli et al. (1995a), we use more informative hyperpriors for the variance parameters, which will help with the MCMC convergence in the case of sparse data.

We performed small experiments to evaluate the sensitivity of the model fitting to the choice of hyperprior parameters using all three datasets studied here by also trying a set of less informative hyperprior parameters with $a_1 = a_2 = 1, b_1 = 0.01, b_2 = 0.1$. We indeed found the results to be insensitive to the choice of hyperprior parameters for the New England dataset, the largest dataset among the three, whereas the more informative hypepriors yield slightly better prediction results than the less informative ones for the smaller datasets of Florida and California. Choice of hyperpriors did not, however, affect the comparison between models. Bernardinelli et al. (1995a)

also found that a more informative prior is preferred with smaller datasets and more discussion on the choice of hyperprior parameters is found in Bernardinelli et al. (1995a) and Best et al. (1999).

### 3.3.4 Prediction

Given the fitted model, it is straightforward to make predictions for the following year. The prediction, $\widehat{Z}_{i,t+1}$, is obtained by sampling from the posterior predictive distribution through forward sampling. For example, with Model (3.3.3) we have for each iteration of the MCMC algorithm,

$$\widehat{Z}_{i,t+1} = X_{i,t}^T\widehat{\boldsymbol{\beta}} + \widehat{\rho}_i(Z_{i,t} - X_{i,t-1}^T\widehat{\boldsymbol{\beta}}) + \widehat{\phi}_i + \widehat{\epsilon}_{i,t+1},\ \widehat{\epsilon}_{i,t+1} \sim N(0, \widehat{\sigma}^2 q_{i,t}).$$

Note that for year $t+1$ we do not expect to know $q_{i,t+1}$ so we assume $q_{i,t+1} = q_{i,t}$, and hence $\psi_{i,t} = 1$ due to the relatively slow change over time of $q_{i,t}$ at each county. This can be further seen in Figure A.5 of the Appendix. For Models 3, 5, and 6, which include random temporal and/or spatio-temporal random effects, we additionally use forward sampling to sample $\widehat{\alpha}_{t+1}$ and $\widehat{\delta}_{i,t+1}$. For instance, predictions for Model 6 have the form:

$$\widehat{Z}_{i,t+1} = X_{i,t}^T\widehat{\boldsymbol{\beta}} + \widehat{\alpha}_{t+1} + \widehat{\phi}_i + \widehat{\delta}_{i,t+1} + \widehat{\epsilon}_{i,t+1},$$

where $\widehat{\epsilon}_{i,t+1} \ \widetilde{N}\ (0, \widehat{\sigma}^2/n_{i,t})$, $\widehat{\alpha}_{t+1} = \widehat{\rho}_\alpha\widehat{\alpha}_t + \widehat{\epsilon}^{\alpha}_{t+1}$ where $\widehat{\epsilon}^{\alpha}_{t+1} \overset{iid}{\sim} N(0, \widehat{\tau}^2_\alpha(1-\widehat{\rho}^2_\alpha))$, and $\widehat{\delta}_{i,t+1} = \widehat{\rho}_\delta\delta_{i,t} + \widehat{\epsilon}^{\delta}_{i,t+1}$ where $(\widehat{\epsilon}^{\delta}_{1,t+1}, \ldots, \widehat{\epsilon}^{\delta}_{n,t+1})^{\mathrm{T}} \sim N(0, \widehat{\tau}^2_\delta(1-\widehat{\rho}^2_\delta)[(1-\widehat{\lambda}_\delta)I_n + \widehat{\lambda}_\delta R]^{-1})$. Finally, posterior predictions in the original scale are obtained by adding back the detrended global mean of $\widehat{Z}_{i,t+1}$ and then taking the exponential.

## 3.4 Application to HIV data

For each HIV dataset of Florida, California and New England states as described in Section 3.2, we first compare Models (3.3.2) and (3.3.3) to all eight models in Section 3.3.2 by withholding the new HIV diagnosis rates in 2014 as testing data and using the rates from 2008-2013 as training data. We then apply Model (3.3.2) to predict the new HIV diagnosis rates in 2015 based on the rates observed from 2008-2014.

### 3.4.1 Model assessment

We employ the mean squared prediction error (MSPE), the Gneiting and Raftery (2007) adaptation of the predictive model choice criterion (PMCC) of Gelfand and Ghosh (1998) and the continuous rank probability score (CRPS) (Gneiting and Raftery, 2007) to evaluate the prediction performance. Let $\boldsymbol{y}_{obs}$ be the 2008-2013 observed HIV and $y_k, k = 1, \ldots, p$ be the observed HIV in 2014 where $p = 47, 33$ and 112 for Florida, California and New England states, respectively and $M$ is the length of the sampling chain. Letting $Y_k, k = 1, \ldots, p$ denote the random variable of the HIV in 2014, we have

$$MSPE = \frac{1}{p}\sum_{k=1}^{p}(E[Y_k|\boldsymbol{y}_{obs}, \hat{\theta}] - y_k)^2,$$

$$PMCC = \sum_{k=1}^{p}\frac{(E[Y_k|\boldsymbol{y}_{obs}, \hat{\theta}] - y_k)^2}{Var[Y_k|\boldsymbol{y}_{obs}, \hat{\theta}]} + \sum_{k=1}^{p}\log Var[Y_k|\boldsymbol{y}_{obs}, \hat{\theta}],$$

and

$$CRPS = \frac{1}{p}\sum_{k=1}^{p}(E_F|Y_k - y_k| - \frac{1}{2}E_F|Y_k - Y_k'|),$$

where $Y_k$ and $Y_k'$ in the CRPS are independent replicates sampled from the posterior predictive distribution $F$. The latter two measures directly take advantage of the

Table 3.1: Prediction results for the log transformed and detrended new HIV diagnosis rates in 2014 for Florida, California, and New England.

| | Florida | | | | California | | | | New England | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | MSPE | PMCC | CRPS | Coverage | MSPE | PMCC | CRPS | Coverage | MSPE | PMCC | CRPS | Coverage |
| (3.3.2) | **0.1098** | **-61.95** | **0.1829** | 0.9574 | **0.0650** | **-58.26** | **0.1357** | 0.9697 | 0.0915 | -156.3 | 0.1686 | 0.9196 |
| (3.3.3) | 0.1320 | -55.92 | 0.1967 | 0.9574 | 0.0691 | -58.18 | 0.1420 | 0.9697 | 0.0912 | -140.6 | 0.1719 | 0.9107 |
| 1 | 0.1274 | -47.73 | 0.2035 | 0.9787 | 0.1034 | -41.50 | 0.1753 | 0.9697 | 0.1209 | -126.6 | 0.1932 | 0.9821 |
| 2 | **0.1088** | -48.67 | 0.1916 | 1.000 | 0.0791 | -30.95 | 0.1804 | 1.000 | 0.0876 | -151.0 | 0.1661 | 0.9643 |
| 3 | 0.1310 | -47.91 | 0.2048 | 0.9362 | 0.1056 | -44.26 | 0.1731 | 0.9091 | 0.1192 | -124.2 | 0.1913 | 0.9018 |
| 4 | 0.1245 | -20.08 | 0.2197 | 1.000 | 0.0983 | -35.43 | 0.1790 | 0.9697 | 0.1140 | -125.6 | 0.1891 | 0.9821 |
| 5 | 0.1123 | -55.10 | 0.1883 | 0.9787 | 0.0768 | -44.23 | 0.1611 | 1.000 | **0.0840** | **-163.5** | **0.1583** | 0.9018 |
| 6 | 0.1172 | -26.88 | 0.1966 | 1.000 | 0.0803 | -43.57 | 0.1572 | 0.9697 | 0.0928 | -152.1 | 0.1669 | 0.9464 |
| 7 | 0.1169 | 208.1 | 0.2215 | 0.4894 | 0.0894 | 12.43 | 0.1643 | 0.7273 | 0.1040 | 1709.5 | 0.2093 | 0.3482 |
| 8 | 0.1171 | -57.92 | 0.1888 | 0.9574 | 0.0723 | -51.11 | 0.1495 | 0.9697 | 0.0870 | -157.4 | 0.1657 | 0.9196 |
| $Z^*_{t-1}$ | 0.1414 | – | – | – | 0.0949 | – | – | – | 0.1187 | – | – | – |

$Z^*_{t-1}$ indicates the most recent year available for each county.

posterior predictive distribution resulting from MCMC sampling. The PMCC is a balanced loss function composed of two parts; the first part is interpreted as a measure of the goodness of fit and the second as a predictive variance penalty term. CRPS assesses how concentrated the predictive distribution is around the observed values. As opposed to MSPE and PMCC measures, the CRPS evaluates the models based on the entire predictive distribution rather than only its first two moments. Models which minimize these measures are preferred. In addition to these three measures, we also compared the empircal coverage with a 95% nominal coverage of all models as well as compared our predictions to the HIV rate reported the previous year for each county (or most recent year available if previous is missing).

Table 3.1 reports the prediction results in the log transformed scale for the 2014 new HIV diagnosis rates using all 10 models over all three regions. Model (3.3.2) stands out among the 10 models in terms of the three prediction performance measures for Florida and California and is the second best for New England. For the California data, Model (3.3.2) uniformly performs best across all four measures. For the Florida data, Model (3.3.2) uniformly minimizes PMCC and CRPS and is comparable in MSPEs to Model 2, which reports the best MSPE but only marginally. For the New

Table 3.2: Prediction results for the original scale new HIV diagnosis rates in 2014 for Florida, California, and New England.

| | Florida | | | | California | | | | New England | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Model | MSPE | PMCC | CRPS | coverage | MSPE | PMCC | CRPS | coverage | MSPE | PMCC | CRPS | coverage |
| (3.3.2) | 75.30 | **221.8** | **3.905** | 0.9574 | **7.231** | **94.84** | **1.466** | 0.9697 | **45.88** | 393.7 | **2.557** | 0.9196 |
| (3.3.3) | **56.89** | 224.7 | **3.906** | 0.9574 | 9.715 | 98.11 | 1.621 | 0.9697 | 66.43 | 399.3 | 2.949 | 0.9107 |
| 1 | 78.26 | 242.9 | 4.387 | 0.9787 | 10.42 | 113.9 | 1.843 | 0.9697 | 67.47 | 442.7 | 3.041 | 0.9821 |
| 2 | 58.04 | 242.0 | 4.152 | 1.0000 | 8.397 | 138.6 | 2.092 | 1.0000 | 67.47 | 416.3 | 2.751 | 0.9643 |
| 3 | 80.29 | 245.1 | 4.367 | 0.9362 | 10.28 | 106.4 | 1.748 | 0.9091 | 61.41 | 442.2 | 2.983 | 0.9018 |
| 4 | 78.37 | 951.5 | 220.5 | 1.0000 | 10.27 | 195.4 | 1.943 | 0.9697 | 65.11 | 471.0 | 3.006 | 0.9821 |
| 5 | 60.62 | 231.7 | 3.990 | 0.9787 | 7.951 | 116.2 | 1.801 | 1.0000 | 55.10 | **385.3** | 2.605 | 0.9018 |
| 6 | 68.08 | 1994.4 | $1.4 \times 10^{30}$ | 1.000 | 8.248 | 585.4 | $9.80 \times 10^{5}$ | 0.9697 | 63.59 | 913.5 | 2.733 | 0.9464 |
| 7 | 76.67 | 531.2 | 4.570 | 0.4894 | 9.255 | 176.3 | 1.678 | 0.7273 | 58.73 | $2.44 \times 10^{3}$ | 3.185 | 0.3482 |
| 8 | 65.67 | 226.8 | 3.922 | 0.9574 | 8.289 | 107.9 | 1.646 | 0.9697 | 57.78 | 398.4 | 2.717 | 0.9196 |
| $Y^*_{t-1}$ | 61.51 | – | – | – | 10.03 | – | – | – | 62.58 | – | – | – |

$Y^*_{t-1}$ indicates the most recent year available for each county.

England states, Model 5 performs best across all measures. Model (3.3.2) corresponds to the third smallest PMCC which is only about 4.6% larger than the minimum and both models (3.3.2) and (3.3.3) report the third smallest MSPEs. Although for this data Model 2 is also highly competitive, the PMCC and CRPS for this model tends to be among the largest for all three datasets making it undesirable for model selection.

Table 3.2 reports the prediction results in the original scale for the 2014 new HIV diagnosis rates using all 10 models over all three regions. Predictions in the original scale are made by taking the posterior median of the sampling chain of $Y_{i,t+1}$. In this case, model (3.3.2) easily stands out among the 10 models in terms of CRPS for all three datasets. Model (3.3.2) also outperforms all others in terms of PMMC for Florida and California and is comparable to the best for New England. In terms of MSPE, we again see that (3.3.2) is selected as the best model for California and New England while (3.3.3) is selected for Florida. Since the ultimate goal of these models is to predict the HIV rate in cases per 100,000 for next year, these results in the original scale are more useful for model selection. Table 3.2 shows that the SVAR model framework is best overall.

It is interesting to note that for the transformed results in Table 3.1, all models

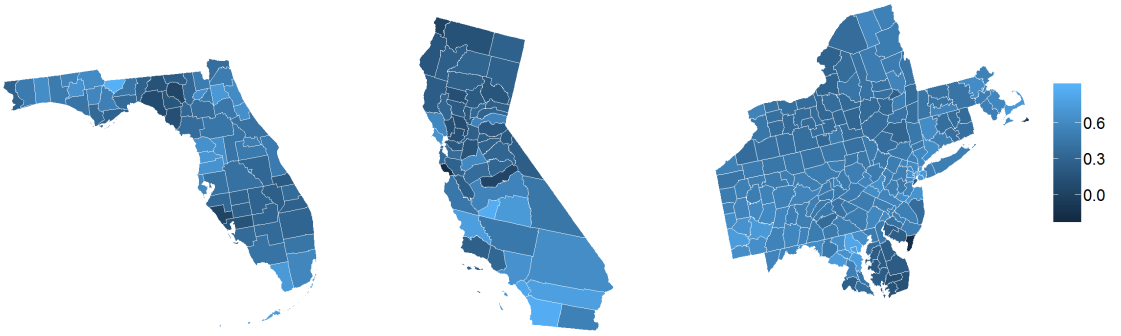


Figure 3.2: Spatial maps of posterior means of $\rho_i$ from model (3.3.2) for Florida, California and New England.

outperform using only the most recently available year as a prediction for 2014 but this is not the case for the results in the original scale of Table 3.2. Although using the most recently available year does not give the best MSPE, it performs better than 70%, 30% and 50% of the models for Florida, California and New England respectively. This inconsistency is not surprising as exponentiation is not a linear function and large posterior sampling values can be very sensitive to the exponentiation.

Across all three regions in Table 3.1, Models 1, 3 and 7 consistently perform among the worst, especially in terms of MSPE and CRPS. Since the unique common feature for those three models is not allowing for spatial correlation, their poor performance implies that incorporating spatial dependence into the model is crucial for making accurate predictions. Furthermore, the fashion in which the spatial dependence is introduced within the model also clearly makes a difference in prediction performance. Our SVAR models seem to carry better prediction power by incorporating the spatial dependence into the varying autoregressive coefficients, compared to the models that simply incorporate the structured spatial random effects in the additive fashion. The significant advantage of allowing for varying autoregressive coefficients is more clearly

shown by comparing the SVAR models to Models 7 and 8 which have invariant autoregressive coefficients. Even though Model 8 includes the structured spatial random effects $\phi_i$, it performs much worse in general than Model (3.3.2) across all measures for Florida and California. Figure 3.2 shows maps of the posterior means of $\rho_i$ from model (3.3.2) for each of the three regions. Apparently, $\rho_i$ varies greatly across each region, which verifies the necessity of our SVAR model for FLorida and California. In the case of New England, Model 8 outperforms Models (3.3.2) and (3.3.3) across all measures. This may indicate that the estimated $\rho_i$s are similar enough that a constant $\rho$ may be appropriaate to capture the autoregressive trend in this case. In fact, the posterior mean estimate for $\lambda_\rho$ is 0.97 which indicates very high spatial correlation among the $\rho_i$s. We can see this homogeneity among the $\rho_i$ for New England in Figure 3.2.

To illustrate the predictions from different models, Figure 3.3 shows the observed and posterior predictive means of $\widehat{Z}_{i,t}$ based on Models (3.3.2), 6 and 7 for the Florida data. The figure also shows the variability of the predictions from these three models. Although county predictions from each model are barely visually discernible, prediction variance for Model (3.3.2) is clearly smaller than that of Model 6. This agrees with the PMCC measures calculated for each of the three models. This result can also be seen in the original scale in Figure 3.4. This figure shows that the 95% credible interval of the posterior sampling distribution for $Y_{i,t+1}$ is worst for Model 6. Although it is also clear in both figures that Model 7 gives the smallest prediction variance, Model 7 gives a significantly smaller empirical coverage than other models which is undesirable. To further illustrate the role of the spatially varying autoregressive function in the prediction of HIV, Figure 3.5 shows a breakdown of the contribution from each of the individual terms, $X\boldsymbol{\beta}$ and $\rho_i(Z_{i,t-1} - X_{i,t-2}\beta)$, in model (3.3.2) by plotting their respective posterior means based on the California data. California was

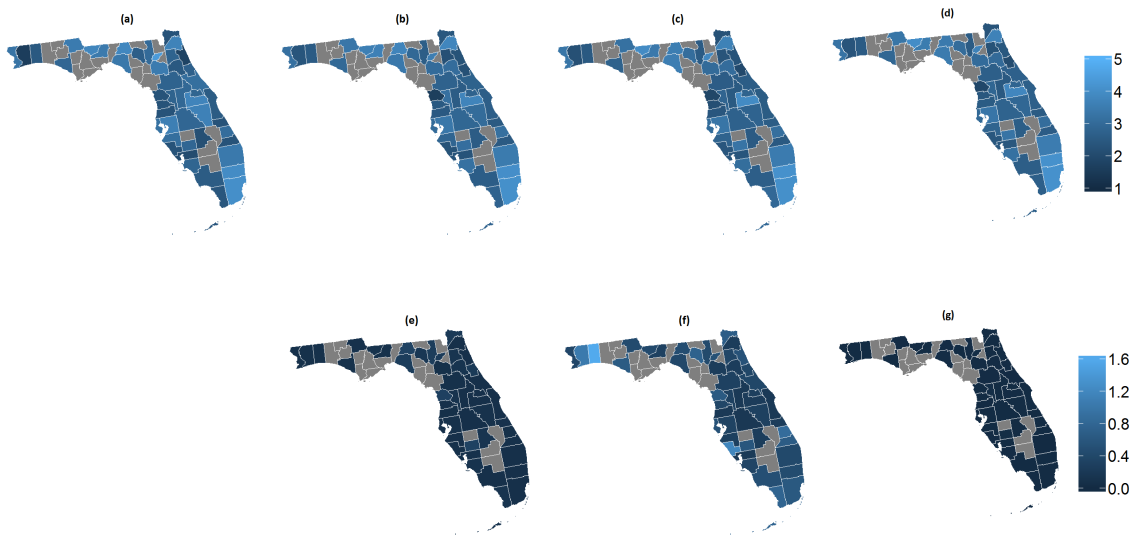


Figure 3.3: Observed (a) and predicted values of $Z_{i,t}$ for 2014 new HIV diagnosis rates in Florida. The predicted values shown in (b)-(d) corresponding prediction variances shown in (e)-(g) are from models (3.3.2), 6 and 7, respectively.


Figure 3.4: Observed (a) and predicted values of $Y_{i,t}$ for 2014 new HIV diagnosis rates in Florida. The predicted values shown in (b)-(d) corresponding widths of the 95% credible intervals shown in (e)-(g) are from models (3.3.2), 6 and 7, respectively.

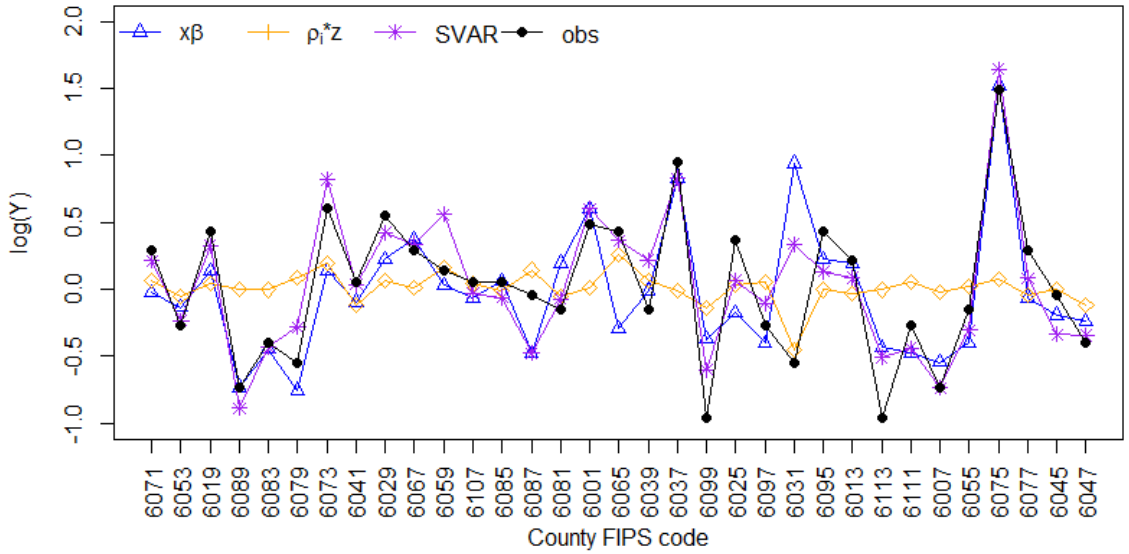


Figure 3.5: Break down of the contribution from each term in Model (3.3.2) to the prediction of 2014 new HIV diagnosis rates in California, where $\rho_i^* Z$ indicates the contribution from $\rho_i(Z_{i,t-1} - X_{i,t-2}\beta)$ and SVAR indicates the overall model prediction.

chosen for illustration because this region has the smallest number of counties and thus provides a better visual. The highly varying pattern of the component attributed to $\rho_i(Z_{i,t-1} - X_{i,t-2}\beta)$ shows that the spatially varying autoregressive function makes an important contribution in capturing the variability of the HIV data.

### 3.4.2 Prediction of 2015 new HIV diagnoses

Due to the sluggish procedure of the public reporting of new HIV diagnosis data, the 2014 HIV data was not released until July 2016. Thus it will take at least until July 2017 to access the 2015 HIV data. We make predictions for the 2015 new HIV diagnoses using Model (3.3.2) based on the the annual HIV data collected from 2008 to 2014 and the available annual demographic information between 2007 to 2014. The website *healthindicators.gov* where we retrieve many of the demographic variables has paused their data collection service in June 2016 due to budget cuts,

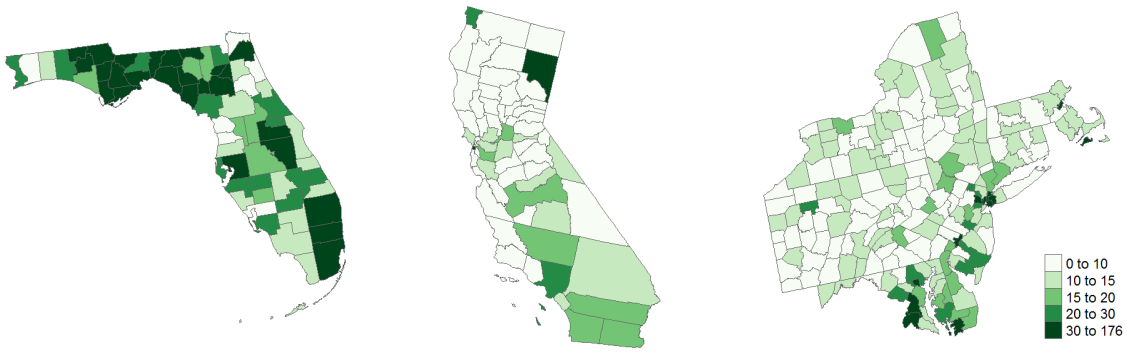


Figure 3.6: Predicted 2015 new HIV diagnosis rates, $Y_{i,t}$, in Florida, California and New England.

having only collected data through 2013, and it is unclear when they would resume the service. This leads to incomplete demographic information for 2014. We used 2014 demographic variables whenever they are available and an average of previous years data when they are missing. The influence of the missing variables should be minimal. On the one hand, we find that the demographic variables are relatively stable over time and on the other hand, only four variables of 2014, two pertaining to weather, excessive drinking and inactivity, are missing from our selected set of 28 demographic variables. Our predictions can provide a timely surrogate for the true data, and thus are crucial for identifying HIV hotspots and making decisions on prevention, testing and treatment efforts.

Figure 3.6 shows the predicted 2015 new HIV diagnosis rates on their original scale. The plot shows that the counties expected to have the highest rates $Y_{i,t}$ in 2015 include: Miami-Dade, FL(175.6), Broward, MD(157.6), Union, FL (94.8), Orange, FL (84.1), Baltimore, MD (78.2), Lassen, CA (63.4), and Prince George's, MD (61.4). The numbers in parentheses are our predicted new HIV diagnosis rates for that county. We find that these counties tend to have a higher percentages of African Americans, people in the age range of 25-44, and people living in same sex households than other

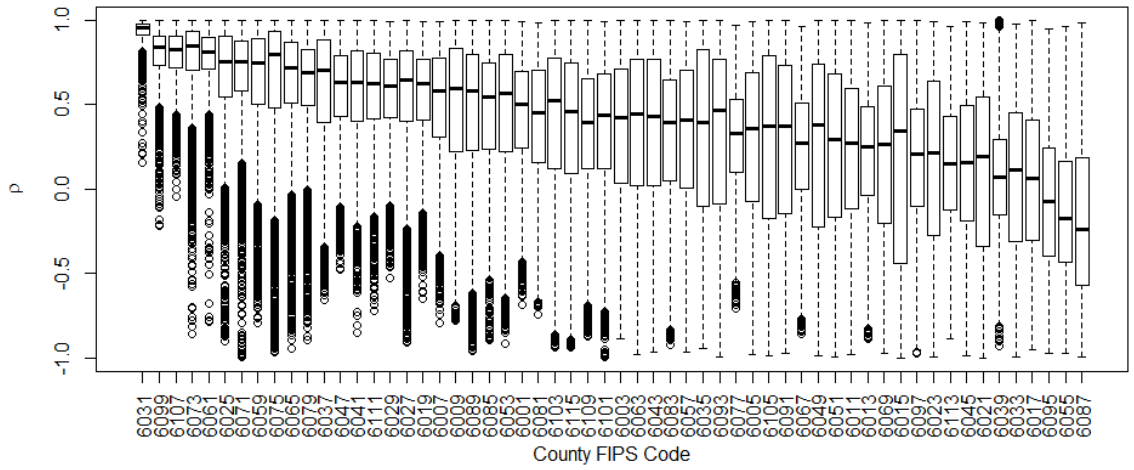


Figure 3.7: Boxplots of posterior samples of $\rho_i$ used for prediction of 2015 new HIV diagnosis rates for all counties in California. The counties are ordered according to the posterior means of $\rho_i$ in a descending order.

counties. It is also seen that Florida in general has higher predicted new diagnosis rates than the other two regions, which is an indication of severe HIV infection in the South.

The SVAR model allows us to study the evolution of HIV over time. Because the demographic variables are relatively steady, a positive autoregressive coefficient indicates that if a high diagnosis rate was observed in the previous year, the following year will likely see an increase in HIV rate. A negative autoregressive coefficient indicates evolution in the opposite direction. The posterior means of most autoregressive coefficients for the counties in the three studied regions are positive. For instance, the boxplot of $\rho_i$ for counties in California, see Figure 3.7, shows that 95% of them are positive. In particular, counties with the largest posterior means of $\rho_i$ can be seen in Figure 3.8. These include most counties in Florida, particularly Miami-Dade, Broward and Hillsborough counties ($> 0.99$), since $\lambda_\rho = 0.97$ indicates strong spatial correlation estimated using Model 3.3.2. Counties in California and New England highest posterior mean estimates for $\rho_i$ in their region include Kings, CA (0.936) and

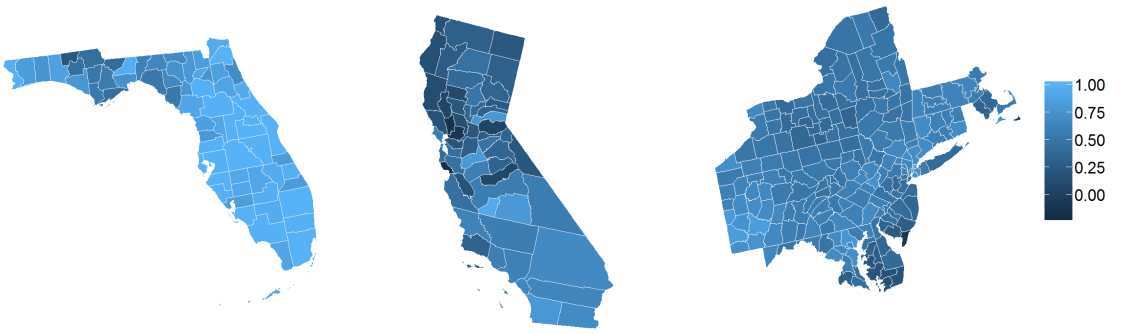


Figure 3.8: Spatial maps of posterior means of $\rho_i$ from model (3.3.2) for Florida, California and New England for the prediction of 2015.

Washington, PA (0.836). Counties with smallest values among all three regions include Santa Cruz, Napa, and Solano, CA with $\rho_i$ estimates between -0.111 and -0.156 as well as Cape May, NJ (-0.071) and Nantucket, MA (-0.008). All estimates of $\rho_i$ for counties in Florida are greater than that of Holmes county (0.212). A caveat for these results is that there are no observations in Nantucket between 2008 and 2014 so this should be taken into account when interpreting the estimated $\rho_i$ for this county. Note that the counties with highest estimates of $\rho_i$ do not necessarily coincide with those having the highest predicted rates in 2015. This is because $\rho_i$ only offers an indicator on how the current observation will develop one year ahead. It is also important to recognize that if counties with missing data are clustered, the interpretation of estimated $\rho_i$ and future predictions for these counties should be made with caution. For example, Lassen County, California is predicted to have a high rate in 2015 but is has no available data from 2008-2014 and 3 of its 4 neighboring counties also have no reported rates during this time frame.

## 3.5 Discussion

For spatially aggregated data with fixed number and location of areas, it is natural to investigate the characteristics of each individual time series when making forecast, so having an individual autoregressive model for each county immediately follows. However, due to the spatial correlation of the time series at adjacent counties and the need to borrow strength from neighbors in model fitting owing to the short length of each series, it is required that the autoregression coefficients follow a Markov spatial random process. For these reasons, we proposed our SVAR models which show great advantages in terms of prediction over the linear mixed models that have been popular for spatio-temporal disease data modeling. In our models, the space-time interaction is incorporated in a unique fashion by allowing for spatially varying autoregressive coefficients. This procedure is more convenient than including a nonseparable spatio-temporally correlated random effect in the model as the nonseparable space-time covariance function is often difficult to model and to manipulate in the computation.

The feature of our SVAR models that allows the temporal autocorrelation coefficients to spatially vary while still remaining parsimonious is particularly important for the HIV datasets analyzed in this article, which have very short time series available for each county. We compared our models to eight competing models and found that our class performs best when jointly viewing their MSPE, PPL and CRPS based on new HIV diagnoses rates of Florida, California and New England. Such SVAR models are also more intuitive to interpret as $\rho_i$ directly measures how HIV, given demographic variables, evolves over time, that is, how the value of next year depends on previous years. In the geostatistics setting, Nobre et al. (2011) proposed a spatially varying autoregressive model with order $p$ to model space-time continuous processes of sea surface temperatures, where kernel convolution of bounded variables,

in particularly beta random variables, are proposed as priors for the bounded process of autoregression coefficients. Our copula modeling for prior of autoregressive coefficients is more computationally efficient than the kernel convolution and can be generalized to any non-Gaussian prior.

Although we find that model (3.3.2) performs the best or among the best with the HIV data from the three regions, this basic form of SVAR model may not outperform its variations, such as model (3.3.3), given another data set or even the same data set but simply over other regions due to varying strengths in association between the disease and the demographic variables and varying dependency structures within different regions of the country. If the demographic variables over one region can explain the spatial/temporal/spatio-temporal variability better than another region, then the spatial/temporal/spatio-temporal structure in this region becomes less important. In general, there is no uniformly best model for all data sets. Final model selection depends on the underlying structure of the data.

There are many ways to expand our SVAR models to make them more flexible to better accommodate different spatio-temporal processes. For example, the AR(1) model can be replaced by a higher order autoregressive model where coefficients at each order are allowed to vary spatially. We take a log transformation to attain the Gaussian properties of our data. However, we are aware that such transformation may not be appropriate for other data sets. Our model can easily be adapted to accommodate other types of transformation such as the square root transformation or Freeman-Tukey transformation in the form of $\sqrt{x} + \sqrt{x+1}$ (Waller et al., 1997), or to fit within other frameworks such as specifying $Y_{i,t}$ to follow a Poisson or binomial distribution. In the latter case, we model $Y_{it} \sim Poisson(E_{it} r_{it})$ or $Y_{it} \sim binomial(n_{it}, p_{it})$, where $E_{it}$ is the expected rates and $n_{it}$ the population at risk for the $i$th county at time $t$. In these cases, we then model $log(r_{it})$ or logit$(p_{it})$ in a

similar fashion as $Z_{it}$ in this paper.

Finally, we focus on developing an effective statistical modeling framework for HIV prediction based on the observed data. The autoregressive nature of our models makes it challenging to perform prediction for counties with suppressed data in previous years. Although it is well-known that the missing data not at random can cause bias in the parameter estimation (Little and Rubin, 2014), its effect on the prediction of our modeling framework is not very clear. A study of how to appropriately incorporate information from the missing data to better inform the prediction is currently underway.

# Chapter 4

# Summary and Final Thoughts

The focus of my dissertation research has been methods and applications in spatio-temporal statistics. Although this is quite a niche topic within statistics, I feel I have barely scratched the surface of the capabilities of spatio-temporal analyses. My work in this field thus far can be divided into two distinct chapters as has been done above: the development of a flexible approach to model nonstationary spatio-temporal data, particularly for environmental data, and the development of a class of hierarchical Bayesian prediction models inspired by space-time disease data. Reflecting, it is interesting to note the stark differences between these two bodies of work. The first employs frequentist methods to model geostatistical data while the second models areal data in a strictly Bayesian hierarchical setting. This final chapter will briefly summarize the two projects, both collaborations with my advisor Dr. Bo Li, the latter also a collaboration with Drs. Trevor Park and Dolores Albarracín.

We proposed a robust nonstationary modeling approach for space-time data. Stationarity is a commonly relied upon assumption in the modeling of stochastic processes. A stationary process implies that the process autocorrelation structure does

not vary over the spatial and/or temporal domain. This assumption is not always appropriate when we have complex underlying spatial and temporal structures as often seen in the natural or life sciences. Our nonstationary modeling approach is achieved by estimating a latent expanded space, for each of the spatial and temporal domains, in which the nonstationary process exhibits stationary properties. Since space-time data can become large rather quickly as the number of spatial locations and/or time points increases, we also propose a strategic sampling approach to deal with a large dataset. Both simulations and real data application to streamflow and wind datasets show that modeling nonstationarity in both space and time can improve the predictive performance over stationary covariance models and models which are stationary only in the spatial or time domain. We also used the streamflow dataset to show our approach outperformed the existing nonstationary models of Ma (2002) and Garg et al. (2012). The benefits of our model are that it can easily accommodate nonstationarity in both space and time domains and allows us to take advantage of the plethora of stationary covariance models available. The proposed strategic sampling approach for large data also makes our model appealing since much of available spatio-temporal data is quite large.

In a Bayesian hierarchical setting, we proposed a class of spatially varying autoregressive (SVAR) models compounded with conditional autoregressive (CAR) spatial correlation structures. These models were motivated by an HIV dataset. The rare nature of the disease is a major challenge of modeling HIV data which leads to few to no incidents across much of the domain. This sparsity makes it necessary to borrow strength from neighboring regions in both space and time in order to more accurately estimate the underlying disease risk. Our SVAR models allow the temporal correlation to be location specific while remaining a low rank model, a feature which is important for the publicly available short history HIV data since individual estimations of

an AR(1) models for each county are infeasible. To model the complex dependence between adjacent counties, we resort to the copula approach and flexible CAR formulation. The copula approach is critical here to allow us to model the marginal distributions of the autocorrelation parameters for each county, $\rho_i,\ i = 1, \cdots, n$, separately from the correlation structure imposed on $\{\rho_1, ..., \rho_n\}$. Bayesian hierarchical models are known to be flexible by efficiently incorporating a multitude of moving parts and are particularly-well suited for studying the transmission of disease and making predictions. We applied the proposed models to HIV data over Florida, California and New England states, and compared them to a range of linear mixed models that have been popular for modeling spatio-temporal disease data. The results show that for such data our proposed models outperform the others in terms of prediction. These models should help health organizations make more informative decisions as regional prediction of disease is central to orchestrating appropriate public health responses. I am also hopeful that these models could be useful to areas outside of public health.

I look forward to building upon the strong foundation of knowledge I've gained over the past five years, to continue to dig deeper into what I have yet to learn and to adapt my skill set to new research fields.

# Bibliography

S. Aberg, F. Lindgren, A. Malmberg, J. Holst, and U. Holst. An image warping approach to spatio-temporal modeling. *Environmentrics*, 16(8):833–848, 2005.

D. Albarracín, M. B. Tannenbaum, L. R. Glasman, and A. J. Rothman. Modeling structural, dyadic, and individual factors: the inclusion and exclusion model of hiv related behavior. *AIDS and Behaviors*, 14(S2):239–249, 2010.

R. M. Assunção, I. A. Reis, and C. D. Oliveira. Diffusion and prediction of leishmaniasis in a large metropolitan area in brazil with a bayesian space-time model. *Statistics in Medicine*, 20(15):2319–2335, 2001.

A. B. Balaji, K. E. Bowles, B. C. Le, and G. Paz-Bailey and A M Oster. High hiv incidence and prevalence and associated factors among young msm. *AIDS*, 27(2): 269–278, 2013.

S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive models for large spatial datasets. *Journal of the Royal Statsitical Society: Series B*, 70: 825–848, 2008.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. Bayesian inference. In F. Buena, V. Isham, N. Keiding, T. Louis, R. L. Smith, and H. Tong, editors, *Hierarchical*

*modeling and analysis for spatial data, 2nd edition*, pages 100–107. CRC Press, 2015.

A. Bárdossy. Copula-based geostatistical models for groundwater quality parameters. *Water Resources Research*, 42(11), 2006. ISSN 1944-7973. W11416.

C. Bauer, J. Wakefield, H. Rue, S. Self, Z. Feng, and Y. Wang. Bayesian penalized spline models for the analysis of spatio-temporal count data. *Statistics in Medicine*, 35(11):1848–1865, 2016.

L. Bernardinelli, D. Clayton, and C. Montomoli. Bayesian estiamtes of disease maps: how important are priors? *Statistics in Medicine*, 14:2411–2431, 1995a.

L. Bernardinelli, D. G. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini. Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, 14(21-22):2433–2443, 1995b.

J. Besag, J. york, and A. Mollie. A bayesian image restoration with two applications in spatial statistics. *Ann Inst Stat Math*, 43:1–59, 2011.

N. G. Best, A. A. Richard, A. Thomas, L. A. Waller, and E. Conlon. Bayesian disease models for spatially correlated disease and exposure data. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian spatial modeling*. Oxford Unviersity Press, 1999.

D. Bolin and F. Lingren. Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Annals of Applied Statistics*, 5:532–550, 2011.

L. Bornn, G. Shaddick, and J. V. Zidek. Modeling nonstationary processes through

dimension expansion. *Journal of the American Statistical Association*, 107(497): 281–289, 2012.

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

B. Cai, A. B. Lawson, M. M. Hossain, J. Choi, R. S. Kirby, and J. Liu. Bayesian semi-parametric model with spatially-temporally varying coefficients selection. *Statistics in Medicine*, 32(21):3670–3685, 2014.

I. Choi, B. Li, and X. Wang. Nonparametric estimation of spatial and space-time covariance functions. *Journal of Agriculture, Biological and Environmental Statistics*, 18:611–630, 2013.

N. Cressie. *Statistics for Spatial Data*, chapter Conditionally and simultaneously specified spatial Gaussian models, pages 402–410. Wiley Series in Probability and Mathematical Statistics, 1991.

N. Cressie. *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics, 1993.

N. Cressie and D. M. Hawkins. Estimation of the variogram. *Journal of the International Association for Mathematical Geology*, 12:115–125, 1980.

N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 70(1):209–226, 2008.

D. Damian, P. D. Sampson, and P. Guttorp. Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmentrics*, 12(2):161–178, 2001.

J. Eidsvik, A. O. Finley, S. Banerjee, and H. Rue. Approximate bayesian inference for large spatial datasets using predictive process models. *Computational Statistics and Data Analysis*, 56:1362–1380, 2012.

A. O. Finley, , H. Sang, S. Banerjee, and A. E. Gelfand. Improving the performance of predictive process modeling for large datasets. *Computational Statistics and Data Analysis*, 53:2873–2884, 2009.

T. C. O. Fonesca and M. F. J. Steel. A general class of nonseparable space-time covariance models. *Environmetrics*, 22(2):224–242, 2011.

M. Fuentes. A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, 12:469–483, 2001.

M. Fuentes. Spectral methods for nonstationary spatial processes. *Biometrika*, 89: 197–210, 2002.

S. Garg, A. Singh, and F. Ramos. Learning non-stationary space-time models for environmental monitoring. Proceedings of the 26th AAAI Conference on Artificial Intelligence, 2012.

A. E. Gelfand and S. K. Ghosh. Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85(1):1–11, 1998.

A. Gelman. Prior distributions for varaince parameters in hierarchical models. *Bayesian Anal*, 1:515–533, 2006.

T. Gneiting. Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association*, 97(458):590–600, 2002.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

R. Guhaniyogi, A. O. Finley, S. Banerjee, and A. E. Gelfand. Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics*, 22:997–1007, 2011.

P. Guttorp, W. Meiring, and P. D. Sampson. A space-time analysis of ground-level ozone data. *Environmetrics*, 5(3):241–254, 1994.

D. Higdon. A process-convolution approach to modeling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, 5(2):173–190, 1998.

D. M. Higdon, J. Swall, and J. Kern. Non-stationary spatial modeling. In J. M. B. et al, editor, *Bayesian Statistics*, volume 6, pages 761–768. Oxford University Press, 1999.

H. Huang and N.-J. Hsu. Modeling transport effects on ground-level ozone using a non-stationary space-time model. *Environmetrics*, 15:251–268, 2004.

M. Jun and M. G. Genton. A test for stationarity of spatio-temporal random fields on planar and spherical domains. *Statistica Sinica*, 22:1737–1764, 2012.

E. L. Kang, N. Cressie, and T. Shi. Using temporal variability to improve spatial mapping with application to satellite data. *Canadian Journal of Statistics*, 38(2): 271–289, 2010.

M. Katzfuss. Bayesian nonstationary spatial modeling for very large datasets. *Environmetrics*, 24(3):189–200, 2013.

H. Kazianka and J. Pilz. Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic Environmental Research and Risk Assessment*, 24:661–673, 2010.

Y. Kim, J. Kim, and Y. Kim. Blockwise sparse regression. *Statistica Sinica*, 16: 375–390, 2006.

L. Knorr-Held. Bayesian modeling of inseparable space-time varaition in disease risk. *Statistics in Medicine*, 19:2555–2567, 2000.

C. Lagazio, A. Biggeri, and E. Dreassi. Age-period-cohort models and disease mapping. *Environmetrics*, 14(5):475–490, 2003.

D. Lee. A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, 2:79–89, 2011.

B. G. Leroux, X. Lei, and N. Breslow. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In M. Halloran and D. Berry, editors, *Statistical models in epidemiology, the environment and clinical trials*. New York: Springer-Verlag, 1999.

B. Li, M. G. Genton, and M. Sherman. A nonparametric assessment of properties of space-time covariance functions. *Journal of the American Statistical Association*, 102(478):736–744, 2007.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B*, 73(4):423–498, 2011.

R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2014.

A. López-Quílez and F. Muñoz. Review of spatio-temporal models for disease mapping. 2009.

C. Ma. Spatio-temporal covariance functions generated by mixtures. *Mathematical Geology*, 34(8):965–975, 2002.

Y. C. MacNab. Hierarchical bayesian modeling of spatially correlated health service outcome and utilization rates. *Biometrics*, 59:305–316, 2003.

Y. C. MacNab and C. B. Dean. Autoregressive spatial smoothing and temporal spline smoothing for mapping rates. *Biometrics*, 57(3):949–956, 2001.

Y. C. MacNab and C. B. Dean. Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in Medicine*, 21(3):347–358, 2002.

M. A. Martínez-Beneito, A. López-Quilez, and P. Botella-Rocamora. An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, 27(15): 2874–2889, 2008.

G. Matheron.

L. D. Mercer, J. Wakefield, A. Pantazis, A. M. Lutambi, H. Mansanja, and S. Clark. Space-time smoothing of complex survey data: small area estimation for child mortality. *Annals of Applied Statistics*, 9(4):1889–1905, 2015.

A. Nobre, B. Sansó, and A. Schmidt. Spatially varying autoregressive processes. *Technometrics*, 53:310–321, 2011.

A. A. Nobre, a M Schmidt, and H. F. Lopes. Spatio-temporal models for mapping the incidence of malaria in pará. *Environmetrics*, 16(3):291–304, 2005.

J. Nocedal and S. J. Wright. *Numerical optimization*. Springer Verlag, Berlin, 1999.

D. Nychka, C. Wikle, and A. Royle. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling: An International Journal*, 2:315–331, 2002.

D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. A multiresolution gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015.

C. J. Paciorek and M. J. Schervish. Nonstationary covariance functions for gaussian process regression. *Advances in Neural Information Processing Systems 16*, 16: 273–280, 2004.

O. Perrin and W. Meiring. Nonstationarity in $\mathbb{R}^n$ is second-order stationary in $\mathbb{R}^{2n}$. *Journal of Applied Probability*, 40:815–820, 2003.

O. Perrin and M. Schlather. Can any multivariate gaussian vector be interpreted as a sample from a stationary random process? *Statistics & Probability Letters*, 77: 881–884, 2007.

Q. Ren and S. Banerjee. Hierarchical factor models for large spatially misaligned datasets: A low-rank predictive process approach. *Biometrics*, 69:19–30, 2013.

M. D. Risser and C. Calder. Regression-based covariance functions for nonstationary modeling. *Environmetrics*, 26(4), 2015.

M. D. Ruiz-Medina, R. M. Espejo, M. D. Ugarte, and A. F. Militino. Functional time series analysis of spatio-temporal epidemiological data. *Stochastic Environmental Research and Risk Assessment*, 28(4):943–954, 2014.

P. D. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance. *Journal of the American Statistical Society*, 87(417):108–119, 1992.

O. Schabenberger and C. A. Gotway. *Statistical methods for spatial data analysis*. Chapman and Hall, A CRC Press Company, 2004.

V. Schmid and L. Held. Bayesian extrapolation of space-time trends in cancer registry data. *Biometrics*, 60(4):1034–1042, 2004.

A. M. Schmidt and A. O'Hagan. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B*, 65(3):743–758, 2003.

F. Sigrist, H. R. Künsch, and W. A. Stahel. A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *Annals of Applied Statistics*, 6(4): 1452–1477, 2012.

N. Simon and R. Tibshirani. Standardization and the group lasso penalty. Technical report, Stanford University, 2011.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, 64 (4):583–639, 2002.

M. L. Stein. Space-time covariance functions. *Journal of the American Statistical Association*, 100(469):310–321, 2005.

M. L. Stein. Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8:1–19, 2014.

H. S. Stern and N. Cressie. Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, 19:2377–2397, 2000.

J. R. Stroud, P. Müller, and B. Sansó. Dynamic models for spatio-temporal data. *Journal of the Royal Statistical Society: Series B*, 63(4):673–689, 2001.

D. Sun, R. K. Tsutakawa, H. Kim, and Z. He. Spatio-temporal interaction with disease mapping. *Statistics in Medicine*, 19(15):2015–2035, 2000.

M. D. Ugarte, T. Goicoa, J. Etxeberria, and A. F. Militino. A p-spline anova type model in space-time disease mapping. *Stochastic Environmental Research and Risk Assessment*, 26:835–845, 2012.

L. A. Waller, B. P. Carlin, H. Xia, and A. E. Gelfand. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92:607–617, 1997.

C. K. Wikle. A kernel-based spectral model for non-gaussian spatio-temporal processes. *Statistical Modeling: An International Journal*, 2(4):299–314, 2002.

C. K. Wikle and N. Cressie. A dimension-reduced approach to space-time kalman filtering. *Biometrika*, 86(4):815–829, 1999.

H. Xia and B. P. Carlin. Spatio-temporal models with errors in covariates: mapping ohio lung cancer mortality. *Statistics in Medicine*, 17(18):2025–2043, 1998.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67, 2006.

# Appendix A

# Exploratory Analysis and Derivations for Full Conditional Distributions of Models Presented in Chapter 3

## A.1 Exploratory Analysis

### A.1.1 Normality Assumptions

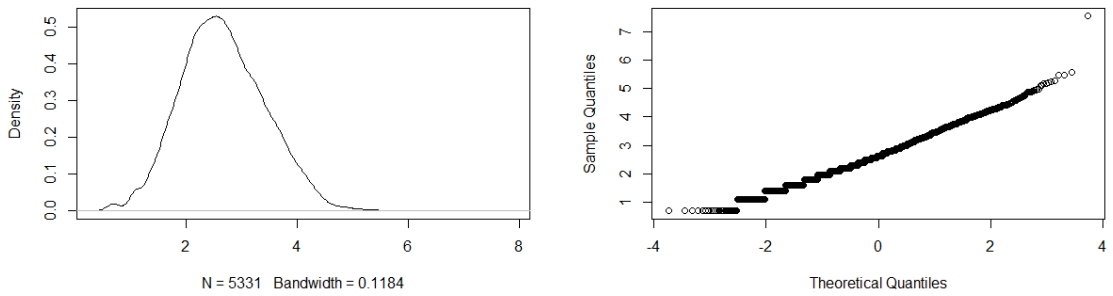


Figure A.1: Density plot (left) and QQ normality plot (right) of $log(Y_{i,t})$ of all available county data in the US from 2008 to 2014.

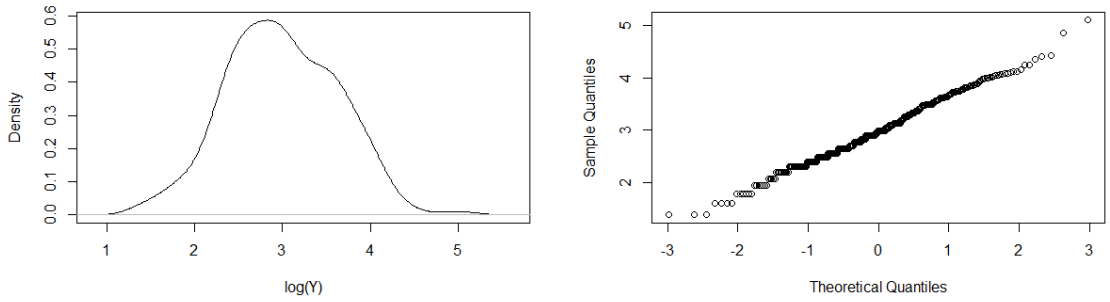


Figure A.2: Density plot (left) and QQ normality plot (right) of $log(Y_{i,t})$ of all available county data in Florida from 2008 to 2014.

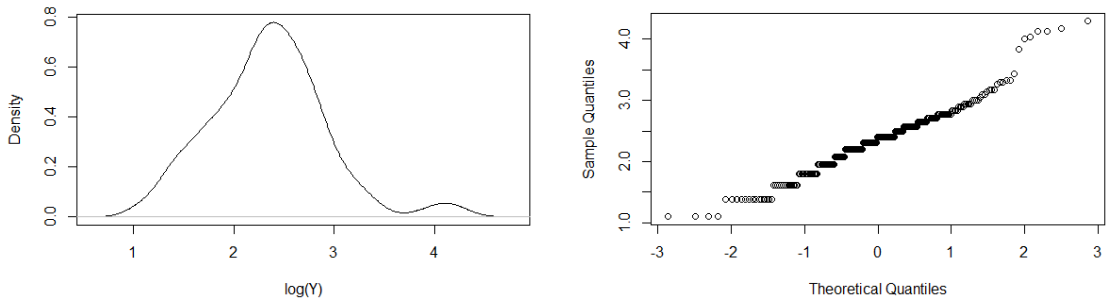


Figure A.3: Density plot (left) and QQ normality plot (right) of $log(Y_{i,t})$ of all available county data in California from 2008 to 2014.

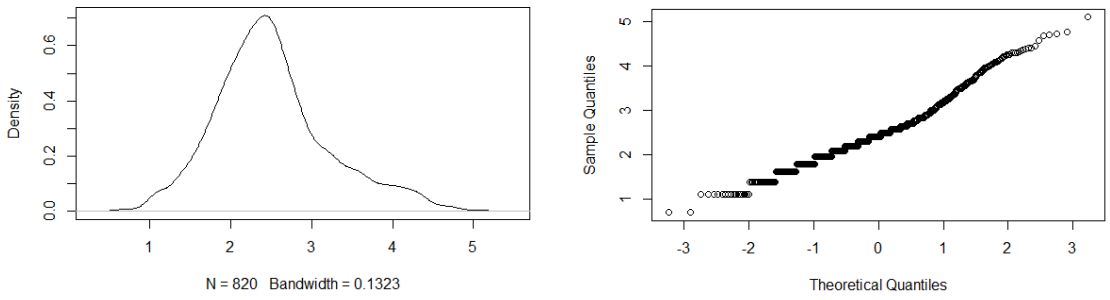


Figure A.4: Density plot (left) and QQ normality plot (right) of $log(Y_{i,t})$ of all available county data in New England from 2008 to 2014.

### A.1.2 Population Size

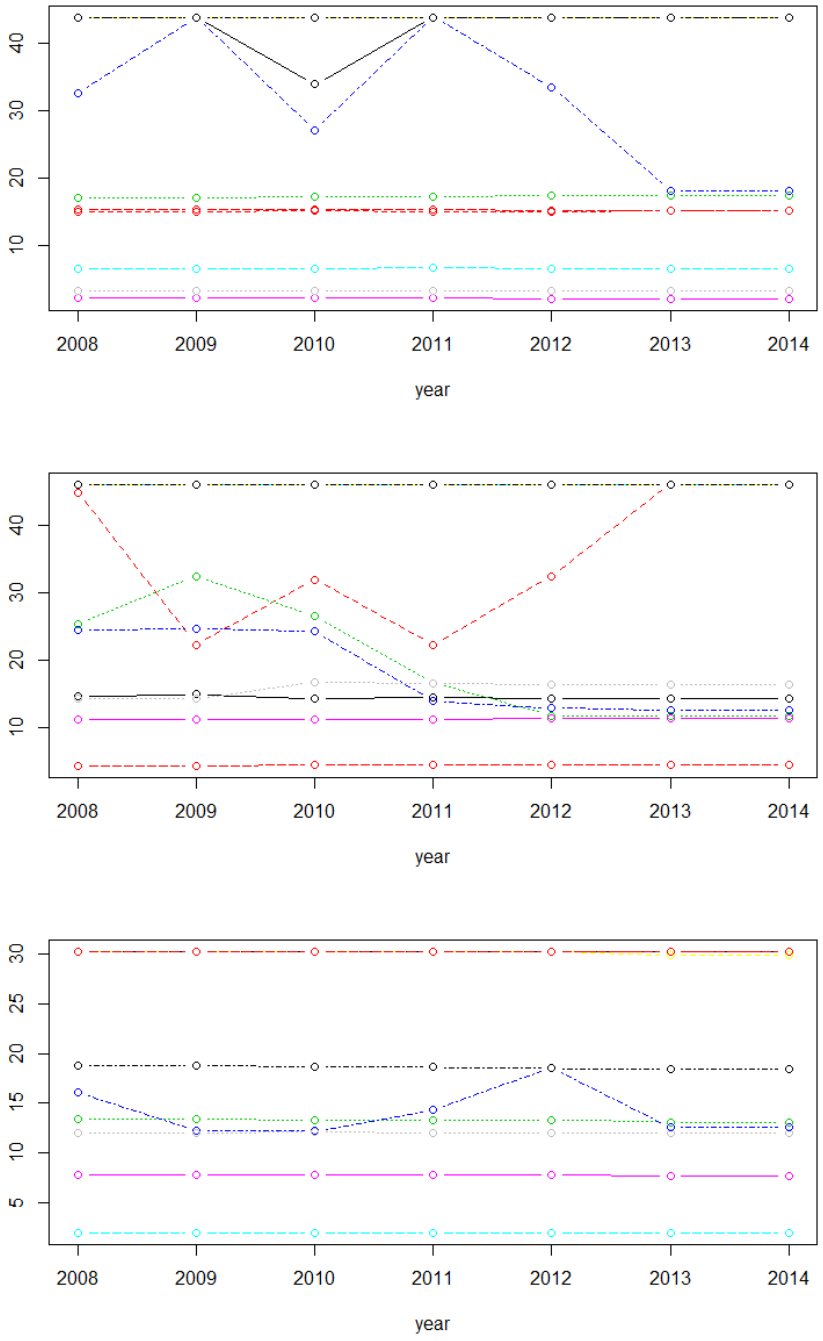


Figure A.5: $q_{i,t} = Y_{i,t} n_{i,t}/c$ of a random selection of 10 counties from Florida (top), California (middle) and North England (bottom).

### A.1.3 Independent AR(1) models

We fit an AR(1) model for each county independently and examined the spatial correlation of the estimated $\rho$s. Low p-values in table A.1 indicate California and New England exhibit positive spatial correlation among the estimated $\rho$s. The independent rho estimates by county (for counties with no more than 3 missing years) can be seen in Figure A.6. Indeed, there is evidence of similar $\rho$ estimates among neighboring counties.

Table A.1: Test Statistics and p-values for for Moran's I and Geary's C testing the null hypothesis of no spatial correlation and the one-sided alternative hypothesis of positive spatial correlation.

| | Florida | | California | | New England | |
|---|---|---|---|---|---|---|
| | Statistic | p-value | Statistic | p-value | Statistic | p-value |
| Moran's I | 0.0343 | 0.3216 | 0.1072 | 0.1389 | 0.2598 | 0.0003 |
| Geary's C | 0.9610 | 0.3758 | 0.8098 | 0.0665 | 0.7384 | 0.0005 |

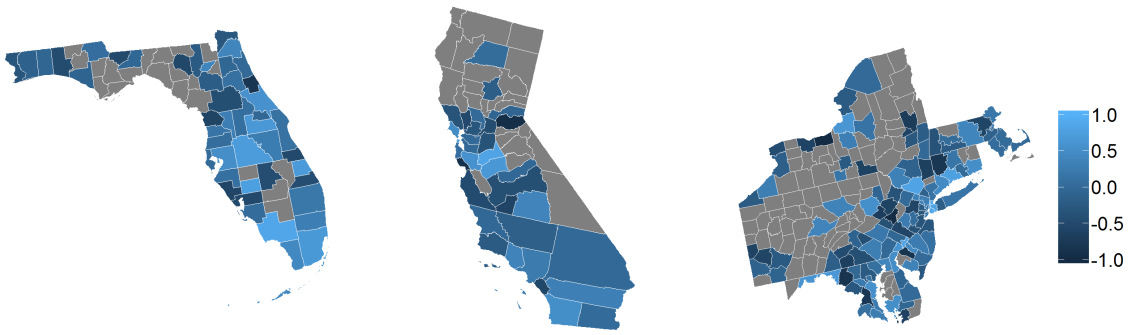


Figure A.6: Spatial maps of independent $\rho$ estimates using maximum likelihood estimation for counties in Florida, California and New England.

## A.2 Full conditional distributions and MH algorithm details.

Define $Z_{i,t} = log(Y_{i,t})$ to be the log transformed new HIV diagnoses rates for $i = 1, \ldots, n,\ t = 1, \ldots, T$ where n and T are the number of available counties and years respectively.

**Normal and Inverse Gamma Distributions**

Normal Distribution: $f(x; \mu, \Sigma) = \frac{1}{2\pi|\Sigma|^{1/2}} exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu))$

Gamma Distribution: $f(x; a, b) = \frac{b^a}{\Gamma(a)} x^{-a-1} exp(-\frac{b}{x})$

### A.2.1 Model 6

Here we derive all full conditional distributions and Metropolis-Hastings (MH) algorithm details for model 6. Full conditionals for all submodels(1-5) follow a similar form.

*Level I.*

$$Z_{i,t} = X_{i,t}\boldsymbol{\beta} + \alpha_t + \phi_i + \delta_{i,t} + \epsilon_{i,t},\ \epsilon \sim N(0, \sigma^2 Q),\quad Q = \text{diag}\big(\frac{c}{Y_{i,t} n_{i,t}}\big)$$

*Level II (Priors).*

$$\boldsymbol{\beta}|\tau_\beta^2 \overset{iid}{\sim} N(0, \tau_\beta^2)$$

$$\boldsymbol{\alpha} \sim N(0, \Sigma_\alpha),\ \boldsymbol{\phi} \sim N(0, \Sigma_\phi),\ \boldsymbol{\delta} \sim N(0, \Sigma_\delta)$$

$$\Sigma_\alpha = \tau_\alpha^2 \rho_\alpha^{|u|},\ \Sigma_\phi = \tau_\phi^2[(1-\lambda_\phi)I + \lambda_\phi R]^{-1},\ \Sigma_\delta = \tau_\delta^2 \rho_\delta^{|u|} \otimes [(1-\lambda_\delta)I + \lambda_\delta R]^{-1}$$

$$\mu_Z = E[\boldsymbol{Z}|\mu, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\delta}] = X\boldsymbol{\beta} + \boldsymbol{\alpha} + \boldsymbol{\phi} + \boldsymbol{\delta},\ \Sigma_Z = \sigma^2 Q$$

*Hyperpriors.*

$\tau_\alpha^2,\ \tau_\phi^2,\ \tau_\delta^2 \sim IG(a_1, b_1),\ \rho_\alpha,\ \rho_\delta \sim Uniform(-1, 1),\ \lambda_\phi,\ \lambda_\delta \sim Uniform[0, 1],\ \sigma^2 \sim IG(a_2, b_2)$

**Model 6 Full conditional distributions**

Full conditional distribution of $\boldsymbol{\beta}$ :

Given prior $\boldsymbol{\beta}|\tau_\beta^2 \overset{iid}{\sim} N(0, \tau_\beta^2)$, $[\boldsymbol{\beta}|\boldsymbol{Z}, \cdot] \propto [\boldsymbol{Z}|\boldsymbol{\beta}, \cdot][\boldsymbol{\beta}|\tau_\beta^2]$

$$\propto exp\big\{-\tfrac{1}{2\sigma^2}(\boldsymbol{Z} - X\boldsymbol{\beta} - \boldsymbol{\alpha} - \boldsymbol{\phi} - \boldsymbol{\delta})' Q^{-1} (\boldsymbol{Z} - X\boldsymbol{\beta} - \boldsymbol{\alpha} - \boldsymbol{\phi} - \boldsymbol{\delta})\big\} \times exp\big\{-\tfrac{1}{2\tau_\beta^2}\boldsymbol{\beta}'\boldsymbol{\beta}\big\}$$

$$\propto exp\big\{-\tfrac{1}{2}[\boldsymbol{\beta}'[-\tfrac{1}{2\sigma^2}X'Q^{-1}X + \tfrac{1}{\tau_\beta^2}]\beta - \beta'[\tfrac{1}{\sigma^2}X'Q^{-1}(\boldsymbol{Z} - \boldsymbol{\alpha} - \boldsymbol{\phi} - \boldsymbol{\delta})]$$
$$-[\tfrac{1}{\sigma^2}(\boldsymbol{Z} - \boldsymbol{\alpha} - \boldsymbol{\phi} - \boldsymbol{\delta})Q^{-1}X]\boldsymbol{\beta}\big\}$$

by completing the square, we have the full conditional

$$\boldsymbol{\beta}|\boldsymbol{Z}, \cdot \sim MVN(\mu_\beta, \Sigma_\beta)$$

$\Sigma_\beta = [\tfrac{1}{\sigma^2}X'Q^{-1}X + \tfrac{1}{\tau_\beta^2}]^{-1}$

$\mu_\beta = \Sigma_\beta \frac{1}{\sigma^2} X' Q^{-1} (\boldsymbol{Z} - \boldsymbol{\alpha} - \boldsymbol{\phi} - \boldsymbol{\delta})$

Full conditional distribution of $\boldsymbol{\alpha}$ :

Given prior $\boldsymbol{\alpha}|\rho_\alpha, \tau_\alpha^2 \sim N(0, \Sigma_\alpha)$, $[\boldsymbol{\alpha}|\boldsymbol{Z}, \cdot] \propto [\boldsymbol{Z}|\boldsymbol{\alpha}, \cdot][\boldsymbol{\alpha}|\tau_\alpha^2, \rho]$

$\propto exp\big\{ -\frac{1}{2\sigma^2}(\boldsymbol{Z} - X\boldsymbol{\beta} - \boldsymbol{\alpha} \otimes 1_n - 1_T \otimes \boldsymbol{\phi} - \boldsymbol{\delta})' Q^{-1} (\boldsymbol{Z} - X\boldsymbol{\beta} - \boldsymbol{\alpha} \otimes 1_n - 1_T \otimes \boldsymbol{\phi} - \boldsymbol{\delta}) \big\}$
$\times exp\big\{ -\frac{1}{2} \boldsymbol{\alpha}' \Sigma_\alpha^{-1} \boldsymbol{\alpha} \big\}$

$\propto exp\big\{ -\frac{1}{2}[\boldsymbol{\alpha}' [\frac{Q_1^{-1}}{\sigma^2} + \Sigma_\alpha^{-1}] \boldsymbol{\alpha} - \boldsymbol{\alpha}' \frac{1}{2\sigma^2} Q_2^{-1} (\boldsymbol{Z} - X\boldsymbol{\beta} - 1_T \otimes \boldsymbol{\phi} - \boldsymbol{\delta}) - \frac{1}{2\sigma^2} (\boldsymbol{Z} - X\boldsymbol{\beta} - 1_T \otimes \boldsymbol{\phi} - \boldsymbol{\delta})^T Q_2^{-1} \boldsymbol{\alpha}] \big\}$

$Q_1$ is a $T \times T$ diagonal matrix with $t^{th}$ diagonal entry as $\Sigma_{i=1}^n \frac{Y_{i,t} n_{i,t}}{c}$ and

$Q_2$ is a concatenation of $n$ $T \times T$ diagonal matrices where the $t^{th}$ diagonal entry of the $i^{th}$ matrix is $\frac{Y_{i,t} n_{i,t}}{c}$.

The full conditional distribution becomes:

$$\boldsymbol{\alpha}|\boldsymbol{Z}, \cdot \sim MVN(\mu_\alpha, \Sigma_\alpha^{**})$$

$\Sigma_\alpha^{**} = [\frac{Q_1^{-1}}{\sigma^2} + \Sigma_\alpha^{-1}]^{-1}$, $\mu_\alpha = \Sigma_\alpha^{**} Q_2^{-1} \frac{1}{\sigma^2} (\boldsymbol{Z} - X\boldsymbol{\beta} - 1_T \otimes \boldsymbol{\phi} - \boldsymbol{\delta})$

Full conditional distribution of $\boldsymbol{\phi}$ :

Given prior $\boldsymbol{\phi}|\lambda_\phi, \tau_\phi^2 \sim N(0, \Sigma_\phi)$, $[\boldsymbol{\phi}|\boldsymbol{Z}, \cdot] \propto [\boldsymbol{Z}|\boldsymbol{\phi}, \cdot][\boldsymbol{\phi}|\tau_\phi^2, \lambda_\phi]$

$$\propto exp\{-\tfrac{1}{2\sigma^2}(\boldsymbol{Z}-X\boldsymbol{\beta}-\boldsymbol{\alpha}\otimes 1_n-1_T\otimes\boldsymbol{\phi}-\boldsymbol{\delta})'Q^{-1}(\boldsymbol{Z}-X\boldsymbol{\beta}-\boldsymbol{\alpha}\otimes 1_n-1_T\otimes\boldsymbol{\phi}-\boldsymbol{\delta})\}$$
$$\times exp\{-\tfrac{1}{2}\boldsymbol{\phi}'\Sigma_\phi^{-1}\boldsymbol{\phi}\}$$

$$\propto exp\{-\tfrac{1}{2}[\boldsymbol{\phi}'[\tfrac{Q_3^{-1}}{\sigma^2}+\Sigma_\phi^{-1}]\boldsymbol{\phi}-\boldsymbol{\phi}'Q_4^{-1}\tfrac{1}{\sigma^2}(\boldsymbol{Z}-X\boldsymbol{\beta}-\boldsymbol{\alpha}\otimes 1_n-\boldsymbol{\delta})$$
$$-\tfrac{1}{\sigma^2}(\boldsymbol{Z}-X\boldsymbol{\beta}-\boldsymbol{\alpha}\otimes 1_n-\boldsymbol{\delta})Q_4^{-1}\boldsymbol{\phi}]\}$$

$Q_3$ is a $n\times n$ diagonal matrix with $i^{th}$ diagonal entry as $\Sigma_{t=1}^{T}\frac{Y_{i,t}n_{i,t}}{c}$ and

$Q_4$ is a concatenation of $T$ $n\times n$ diagonal matrices where the $i^{th}$ diagonal entry of the $t^{th}$ matrix is $\frac{Y_{i,t}n_{i,t}}{c}$.

The full conditional distribution becomes:

$$\boldsymbol{\phi}|\boldsymbol{Z},\cdot\sim MVN(\mu_\phi,\Sigma_\phi^{**})$$

$$\Sigma_\phi^{**}=[\tfrac{Q_3^{-1}}{\sigma^2}+\Sigma_\phi^{-1}]^{-1},\ \mu_\phi=\Sigma_\phi^{**}\tfrac{1}{\sigma^2}Q_4^{-1}(\boldsymbol{Z}-X\boldsymbol{\beta}-\boldsymbol{\alpha}\otimes 1_n-\boldsymbol{\delta})$$

Full conditional distribution of $\boldsymbol{\delta}$ :

Given prior $\boldsymbol{\delta}|\rho_\delta,\lambda_\delta,\tau_\delta^2\sim N(0,\Sigma_\delta)$, $[\boldsymbol{\delta}|\boldsymbol{Z},\cdot]\propto[\boldsymbol{Z}|\boldsymbol{\delta},\cdot][\boldsymbol{\delta}|\tau_\delta^2,\lambda_\delta,\rho_\delta]$

$$\propto exp\{-\tfrac{1}{2\sigma^2}(\boldsymbol{Z}-X\boldsymbol{\beta}-\boldsymbol{\alpha}\otimes 1_n-1_T\otimes\boldsymbol{\phi}-\boldsymbol{\delta})'Q^{-1}(\boldsymbol{Z}-X\boldsymbol{\beta}-\boldsymbol{\alpha}\otimes 1_n-1_T\otimes\boldsymbol{\phi}-\boldsymbol{\delta})\}$$
$$\times exp\{-\tfrac{1}{2}\boldsymbol{\delta}'\Sigma_\delta^{-1}\boldsymbol{\delta}\}$$

$$\propto exp\{-\tfrac{1}{2}[\boldsymbol{\delta}'[\tfrac{1}{\sigma^2}+\Sigma_\delta^{-1}]\boldsymbol{\delta}-\boldsymbol{\delta}'[\tfrac{1}{\sigma^2}Q^{-1}(\boldsymbol{Z}-X\boldsymbol{\beta}-\boldsymbol{\alpha}\otimes 1_n-1_T\otimes\boldsymbol{\phi})]$$

$-\frac{1}{\sigma^2}(\boldsymbol{Z} - X\boldsymbol{\beta} - \boldsymbol{\alpha} \otimes 1_n - 1_T \otimes \boldsymbol{\phi})' Q^{-1} \boldsymbol{\delta}\big\}$

The full conditional distribution becomes:

$$\boldsymbol{\delta}|\boldsymbol{Z}, \cdot \sim MVN(\mu_\delta, \Sigma_\delta^{**})$$

where $\Sigma_\delta^{**} = [\frac{Q^{-1}}{\sigma^2} + \Sigma_\delta^{-1}]^{-1},\ \mu_\delta = \Sigma_\delta^{**} \frac{1}{\sigma^2} Q^{-1}(\boldsymbol{Z} - X\boldsymbol{\beta} - \boldsymbol{\alpha} \otimes 1_n - 1_T \otimes \boldsymbol{\phi})$

Full conditional distribution of $\sigma^2$ :

Given prior $\sigma^2|a_2, b_2 \sim IG(a_2, b_2), \quad [\sigma^2|\boldsymbol{Z}, \cdot] \propto [\boldsymbol{Z}|\sigma^2, \cdot][\sigma^2|a_2, b_2]$

$\propto (\sigma^2)^{-nT/2} exp\big\{ -\frac{1}{2\sigma^2}(\boldsymbol{Z} - \mu_Z)' Q^{-1}(\boldsymbol{Z} - \mu_Z)\big\} \times (\sigma^2)^{-a_2 - 1} exp\big\{ -\frac{b_2}{\sigma^2}\big\}$

$\propto (\sigma^2)^{-(nT/2 + a_2) - 1} exp\big\{ -\frac{1}{\sigma^2}\big[\frac{1}{2}(\boldsymbol{Z} - \mu_Z)' Q^{-1}(\boldsymbol{Z} - \mu_Z) + b_2\big]\big\}$

The full conditional distribution becomes:

$$\sigma^2|\boldsymbol{Z}, \cdot \sim IG\big(\frac{nT}{2} + a_2, \frac{1}{2}(\boldsymbol{Z} - \mu_Z)' Q^{-1}(\boldsymbol{Z} - \mu_Z) + b_2\big)$$

Full conditional distribution of $\tau_\alpha^2$ :

Let $\Sigma_\alpha = \tau_\alpha^2 \Sigma_\alpha^*$

Given prior $\tau_\alpha^2|a_1, b_1 \sim IG(a_1, b_1), \quad [\tau_\alpha^2|\boldsymbol{Z}, \boldsymbol{\alpha}, \cdot] \propto [\boldsymbol{\alpha}|\tau_\alpha^2, \rho_\alpha][\tau_\alpha^2|a_1, b_1]$

$\propto (\tau_\alpha^2)^{-T/2} exp\{-\frac{1}{2\tau_\alpha^2}\boldsymbol{\alpha}'\Sigma_\alpha^*\boldsymbol{\alpha}\} \times (\tau_\alpha^2)^{-a_1-1} exp\big\{-\frac{b_1}{\tau_\alpha^2}\big\}$

$\propto (\tau_\alpha^2)^{-(T/2+a_1)-1} exp\big\{-\frac{1}{\tau_\alpha^2}\big[\frac{1}{2}\boldsymbol{\alpha}'\Sigma_\alpha^*\boldsymbol{\alpha}+b_1\big]\big\}$

The full conditional distribution becomes:

$$\tau_\alpha^2|\boldsymbol{Z},\boldsymbol{\alpha},\cdot \sim \text{IG}\big(\frac{T}{2}+a_1, \frac{1}{2}\boldsymbol{\alpha}'\Sigma_\alpha^*\boldsymbol{\alpha}+b_1\big)$$

Full conditional distribution of $\tau_\phi^2$ :

Let $\Sigma_\phi = \tau_\phi^2\Sigma_\phi^*$

Given prior $\tau_\phi^2|a_1,b_1 \sim IG(a_1,b_1)$, $[\tau_\phi^2|\boldsymbol{Z},\boldsymbol{\phi},\cdot] \propto [\boldsymbol{\phi}|\tau_\phi^2,\lambda_\phi][\tau_\phi^2|a_1,b_1]$

$\propto (\tau_\phi^2)^{-n/2} exp\{-\frac{1}{2\tau_\phi^2}\boldsymbol{\phi}'\Sigma_\phi^*\boldsymbol{\phi}\} \times (\tau_\phi^2)^{-a_1-1} exp\big\{-\frac{b_1}{\tau_\phi^2}\big\}$

$\propto (\tau_\phi^2)^{-(n/2+a_1)-1} exp\big\{-\frac{1}{\tau_\phi^2}\big[\frac{1}{2}\boldsymbol{\phi}'\Sigma_\phi^*\boldsymbol{\phi}+b_1\big]\big\}$

The full conditional distribution becomes:

$$\tau_\phi^2|\boldsymbol{Z},\boldsymbol{\phi},\cdot \sim \text{IG}\big(\frac{n}{2}+a_1, \frac{1}{2}\boldsymbol{\phi}'\Sigma_\phi^*\boldsymbol{\phi}+b_1\big)$$

Full conditional distribution of $\tau_\delta^2$ :

Let $\Sigma_\delta = \tau_\delta^2 \Sigma_\delta^*$

Given prior $\tau_\delta^2 | a_1, b_1 \sim IG(a_1, b_1)$, $[\tau_\delta^2 | \boldsymbol{Z}, \boldsymbol{\delta}, \cdot] \propto [\boldsymbol{\delta} | \tau_\delta^2, \lambda_\delta, \rho_\delta][\tau_\delta^2 | a_1, b_1]$

$\propto (\tau_\delta^2)^{-nT/2} exp\{-\frac{1}{2\tau_\delta^2} \boldsymbol{\delta}' \Sigma_\delta^* \boldsymbol{\delta}\} \times (\tau_\delta^2)^{-a_1 - 1} exp\{ -\frac{b_1}{\tau_\delta^2} \}$

$\propto (\tau_\delta^2)^{-(nT/2 + a_1) - 1} exp\{ -\frac{1}{\tau_\delta^2} \left[ \frac{1}{2} \boldsymbol{\delta}' \Sigma_\delta^* \boldsymbol{\delta} + b_1 \right] \}$

The full conditional distribution becomes:

$$\tau_\delta^2 | \boldsymbol{Z}, \boldsymbol{\delta}, \cdot \sim \text{IG}\big( \frac{nT}{2} + a_1, \frac{1}{2} \boldsymbol{\delta}' \Sigma_\delta^* \boldsymbol{\delta} + b_1 \big)$$

**Model 6 MH sampling algorithms**

Sampling $\rho_\alpha$ at iteration $k$ :

1. Draw $\rho_\alpha^*$ from proposal $q(\rho_\alpha^* | \rho_\alpha^{(k-1)}) \sim N(\rho_\alpha^{(k-1)}, s_\alpha^2) \mathbb{1}\{0 \leq \rho_\alpha^* \leq 1\}$.

   where $q(\rho_\alpha^* | \rho_\alpha^{(k-1)}) = \frac{\phi\left( \frac{\rho_\alpha^* - \rho_\alpha^{(k-1)}}{s_\alpha} \right)}{\Phi\left( \frac{1 - \rho_\alpha^{(k-1)}}{s_\alpha} \right) - \Phi\left( \frac{-\rho_\alpha^{(k-1)}}{s_\alpha} \right)}$.

2. Calculate acceptance ratio $r_\alpha = \frac{p(\rho_\alpha^* | \cdot) q(\rho_\alpha^{(k-1)} | \rho_\alpha^*)}{p(\rho_\alpha^{(k-1)} | \cdot) q(\rho_\alpha^* | \rho_\alpha^{(k-1)})}$

$$\propto \frac{|\Sigma_\alpha(\rho_\alpha^*)|^{-1/2} exp\{-\frac{1}{2}\boldsymbol{\alpha}'\Sigma_\alpha^{-1}(\rho_\alpha^*)\boldsymbol{\alpha}\}\left[\Phi\left(\frac{1-\rho_\alpha^{(k-1)}}{s_\alpha}\right)-\Phi\left(\frac{-1-\rho_\alpha^{(k-1)}}{s_\alpha}\right)\right]}{|\Sigma_\alpha(\rho_\alpha^{(k-1)})|^{-1/2} exp\{-\frac{1}{2}\boldsymbol{\alpha}'\Sigma_\alpha^{-1}(\rho_\alpha^{(k-1)})\boldsymbol{\alpha}\}\left[\Phi\left(\frac{1-\rho_\alpha^*}{s_\alpha}\right)-\Phi\left(\frac{-1-\rho_\alpha^*}{s_\alpha}\right)\right]}$$

$$= \left[\frac{|\Sigma_\alpha(\rho_\alpha^*)|}{|\Sigma_\alpha(\rho_\alpha^{(k-1)})|}\right]^{-1/2} exp\big\{-\frac{1}{2}\boldsymbol{\alpha}'[\Sigma_\alpha^{-1}(\rho_\alpha^*)-\Sigma_\alpha^{-1}(\rho_\alpha^{(k-1)})]\boldsymbol{\alpha}\big\}\left[\frac{\Phi\left(\frac{1-\rho_\alpha^{(k-1)}}{s_\alpha}\right)-\Phi\left(\frac{-1-\rho_\alpha^{(k-1)}}{s_\alpha}\right)}{\Phi\left(\frac{1-\rho_\alpha^*}{s_\alpha}\right)-\Phi\left(\frac{-1-\rho_\alpha^*}{s_\alpha}\right)}\right]$$

3. Generate $\xi \sim Uniform(0,1)$.

$$\rho_\alpha^{(k)} = \begin{cases} \rho_\alpha^* & if\ r_\alpha > \xi \text{ (with acceptance rate} = min\{1, r_\alpha\}), \\ \rho_\alpha^{(k-1)} & o.w. \end{cases}$$

Sampling $\lambda_\phi$ at iteration $k$ :

1. Draw $\lambda_\phi^*$ from proposal $q(\lambda_\phi^*|\lambda_\phi^{(k-1)}) \sim N(\lambda_\phi^{(k-1)}, s_\phi^2)\mathbb{1}\{0 \le \lambda_\phi^* \le 1\}$.

where $q(\lambda_\phi^*|\lambda_\phi^{(k-1)}) = \frac{\phi\left(\frac{\lambda_\phi^*-\lambda_\phi^{(k-1)}}{s_\phi}\right)}{\Phi\left(\frac{1-\lambda_\phi^{(k-1)}}{s_\phi}\right)-\Phi\left(\frac{-\lambda_\phi^{(k-1)}}{s_\phi}\right)}$.

2. Calculate acceptance ratio $r_\phi = \frac{p(\lambda_\phi^*|\cdot)q(\lambda_\phi^{(k-1)}|\lambda_\phi^*)}{p(\lambda_\phi^{(k-1)}|\cdot)q(\lambda_\phi^*|\lambda_\phi^{(k-1)})}$

$$\propto \frac{|\Sigma_\phi(\lambda_\phi^*)|^{-1/2} exp\{-\frac{1}{2}\boldsymbol{\phi}'\Sigma_\phi^{-1}(\lambda_\phi^*)\boldsymbol{\phi}\}\left[\Phi\left(\frac{1-\lambda_\phi^{(k-1)}}{s_\phi}\right)-\Phi\left(\frac{-\lambda_\phi^{(k-1)}}{s_\phi}\right)\right]}{|\Sigma_\phi(\lambda_\phi^{(k-1)})|^{-1/2} exp\{-\frac{1}{2}\boldsymbol{\phi}'\Sigma_\phi^{-1}(\lambda_\phi^{(k-1)})\boldsymbol{\phi}\}\left[\Phi\left(\frac{1-\lambda_\phi^*}{s_\phi}\right)-\Phi\left(\frac{-\lambda_\phi^*}{s_\phi}\right)\right]}$$

$$= \left[\frac{|\Sigma_\phi(\lambda_\phi^*)|}{|\Sigma_\phi(\lambda_\phi^{(k-1)})|}\right]^{-1/2} exp\big\{-\frac{1}{2}\boldsymbol{\phi}'[\Sigma_\phi^{-1}(\lambda_\phi^*)-\Sigma_\phi^{-1}(\lambda_\phi^{(k-1)})]\boldsymbol{\phi}\big\}\left[\frac{\Phi\left(\frac{1-\lambda_\phi^{(k-1)}}{s_\phi}\right)-\Phi\left(\frac{-\lambda_\phi^{(k-1)}}{s_\phi}\right)}{\Phi\left(\frac{1-\lambda_\phi^*}{s_\phi}\right)-\Phi\left(\frac{-\lambda_\phi^*}{s_\phi}\right)}\right]$$

3. Generate $\xi \sim Uniform(0,1)$.

$$\lambda_\phi^{(k)} = \begin{cases} \lambda_\phi^* & if\ r_\phi > \xi \text{ (with acceptance rate} = min\{1, r_\phi\}), \\ \lambda_\phi^{(k-1)} & o.w. \end{cases}$$

Sampling $\rho_\delta$ at iteration $k$ :

1. Draw $\rho_\delta^*$ from proposal $q(\rho_\delta^*|\rho_\delta^{(k-1)}) \sim N(\rho_\delta^{(k-1)}, s_\delta^2)\mathbb{1}\{0 \leq \rho_\delta^* \leq 1\}$.

   where $q(\rho_\delta^*|\rho_\delta^{(k-1)}) = \frac{\phi\left(\frac{\rho_\delta^* - \rho_\delta^{(k-1)}}{s_\delta}\right)}{\Phi\left(\frac{1-\rho_\delta^{(k-1)}}{s_\delta}\right) - \Phi\left(\frac{-\rho_\delta^{(k-1)}}{s_\delta}\right)}$.

2. Calculate acceptance ratio $r_\delta = \frac{p(\rho_\delta^*|\cdot)q(\rho_\delta^{(k-1)}|\rho_\delta^*)}{p(\rho_\delta^{(k-1)}|\cdot)q(\rho_\delta^*|\rho_\delta^{(k-1)})}$

   $$\propto \frac{|\Sigma_\delta(\rho_\delta^*)|^{-1/2} exp\{-\frac{1}{2}\boldsymbol{\delta}'\Sigma_\delta^{-1}(\rho_\delta^*)\boldsymbol{\delta}\}\left[\Phi\left(\frac{1-\rho_\delta^{(k-1)}}{s_\delta}\right) - \Phi\left(\frac{-1-\rho_\delta^{(k-1)}}{s_\delta}\right)\right]}{|\Sigma_\delta(\rho_\delta^{(k-1)})|^{-1/2} exp\{-\frac{1}{2}\boldsymbol{\delta}'\Sigma_\delta^{-1}(\rho_\delta^{(k-1)})\boldsymbol{\delta}\}\left[\Phi\left(\frac{1-\rho_\delta^*}{s_\delta}\right) - \Phi\left(\frac{-1-\rho_\delta^*}{s_\delta}\right)\right]}$$

   $$= \left[\frac{|\Sigma_\delta(\rho_\delta^*)|}{|\Sigma_\delta(\rho_\delta^{(k-1)})|}\right]^{-1/2} exp\Big\{-\frac{1}{2}\boldsymbol{\delta}'[\Sigma_\delta^{-1}(\rho_\delta^*) - \Sigma_\delta^{-1}(\rho_\delta^{(k-1)})]\boldsymbol{\delta}\Big\}\left[\frac{\Phi\left(\frac{1-\rho_\delta^{(k-1)}}{s_\delta}\right) - \Phi\left(\frac{-1-\rho_\delta^{(k-1)}}{s_\delta}\right)}{\Phi\left(\frac{1-\rho_\delta^*}{s_\delta}\right) - \Phi\left(\frac{-1-\rho_\delta^*}{s_\delta}\right)}\right]$$

3. Generate $\xi \sim Uniform(0, 1)$.

$$\rho_\delta^{(k)} = \begin{cases} \rho_\delta^* & if\ r_\delta > \xi \text{ (with acceptance rate} = min\{1, r_\delta\}), \\ \rho_\delta^{(k-1)} & o.w. \end{cases}$$

Sampling $\lambda_\delta$ at iteration $k$ :

1. Draw $\lambda_\delta^*$ from proposal $q(\lambda_\delta^*|\lambda_\delta^{(k-1)}) \sim N(\lambda_\delta^{(k-1)}, s_{\delta 1}^2)\mathbb{1}\{0 \leq \lambda_\delta^* \leq 1\}$.

   where $q(\lambda_\delta^*|\lambda_\delta^{(k-1)}) = \frac{\phi\left(\frac{\lambda_\delta^* - \lambda_\delta^{(k-1)}}{s_{\delta 1}}\right)}{\Phi\left(\frac{1-\lambda_\delta^{(k-1)}}{s_{\delta 1}}\right) - \Phi\left(\frac{-\lambda_\delta^{(k-1)}}{s_{\delta 1}}\right)}$.

2. Calculate acceptance ratio $r_{\delta 1} = \frac{p(\lambda_\delta^*|\cdot)q(\lambda_\delta^{(k-1)}|\lambda_\delta^*)}{p(\lambda_\delta^{(k-1)}|\cdot)q(\lambda_\delta^*|\lambda_\delta^{(k-1)})}$

$$\propto \frac{|\Sigma_\delta(\lambda_\delta^*)|^{-1/2} exp\{-\frac{1}{2}\boldsymbol{\delta}'\Sigma_\delta^{-1}(\lambda_\delta^*)\boldsymbol{\delta}\}\left[\Phi\left(\frac{1-\lambda_\delta^{(k-1)}}{s_{\delta 1}}\right)-\Phi\left(\frac{-\lambda_\delta^{(k-1)}}{s_{\delta 1}}\right)\right]}{|\Sigma_\delta(\lambda_\delta^{(k-1)})|^{-1/2} exp\{-\frac{1}{2}\boldsymbol{\delta}'\Sigma_\delta^{-1}(\lambda_\delta^{(k-1)})\boldsymbol{\delta}\}\left[\Phi\left(\frac{1-\lambda_\delta^*}{s_{\delta 1}}\right)-\Phi\left(\frac{-\lambda_\delta^*}{s_{\delta 1}}\right)\right]}$$

$$= \left[\frac{|\Sigma_\delta(\lambda_\delta^*)|}{|\Sigma_\delta(\lambda_\delta^{(k-1)})|}\right]^{-1/2} exp\Big\{-\frac{1}{2}\boldsymbol{\delta}'[\Sigma_\delta^{-1}(\lambda_\delta^*)-\Sigma_\phi^{-1}(\lambda_\delta^{(k-1)})]\boldsymbol{\delta}\Big\}\left[\frac{\Phi\left(\frac{1-\lambda_\delta^{(k-1)}}{s_{\delta 1}}\right)-\Phi\left(\frac{-\lambda_\delta^{(k-1)}}{s_{\delta 1}}\right)}{\Phi\left(\frac{1-\lambda_\delta^*}{s_{\delta 1}}\right)-\Phi\left(\frac{-\lambda_\delta^*}{s_{\delta 1}}\right)}\right]$$

3. Generate $\xi \sim Uniform(0,1)$.

$$\lambda_\delta^{(k)} = \begin{cases} \lambda_\delta^* & if\ r_{\delta 1} > \xi \text{ (with acceptance rate} = min\{1, r_{\delta 1}\}), \\ \lambda_\delta^{(k-1)} & o.w. \end{cases}$$

### A.2.2 SVAR Model

Here we derive all full conditional distributions and MH sampling algorithm details for the SVAR model. Models 7 and 8 are special cases of this model.

*Level I.*

$$Z_{i,t} = X_{i,t-1}\boldsymbol{\beta} + \psi_{i,t-1}\rho_i(Z_{i,t-1} - X_{i,t-2}\boldsymbol{\beta}) + \epsilon_{i,t}\ \boldsymbol{\epsilon} \overset{iid}{\sim} N(0, \sigma^2 N)$$

$$\mu_Z = E[\boldsymbol{Z}] = X\boldsymbol{\beta},\ Cov(Z_{\cdot,t}, Z_{\cdot,t-u}) = \begin{cases} \frac{c}{\sqrt{Y_{i,t}n_{i,t}Y_{i,t-u}n_{i,t-u}}}\frac{\sigma^2\boldsymbol{\rho}^{|u|}}{1-\boldsymbol{\rho}^2} & h = 0 \\ 0 & o.w. \end{cases}$$

$$f(\boldsymbol{Z}|\boldsymbol{\beta}, \boldsymbol{\rho}, \sigma^2) = (2\pi)^{-nT/2}|\Sigma_Z|^{-1/2} exp\Big\{-\frac{1}{2}(Z-\mu_Z)'\Sigma_Z^{-1}(Z-\mu_Z)\Big\}$$

*Level II (Priors).*

$$\boldsymbol{\beta}|\tau_\beta^2 \overset{iid}{\sim} N(0, \tau_\beta^2)$$

$\rho_i = \Phi(W_i)$ such that $\rho_i \in (-1, 1)$ where $\boldsymbol{W} \sim N(0, \tau_\rho^2[(1-\lambda_\rho)I + \lambda_\rho R]^{-1})$

$\boldsymbol{\rho}$ is defined through the copula density $g(u_1, ..., u_n)$

$= \frac{\phi_{\Sigma_W}(F_1^{-1}(u_1),...,F_n^{-1}(u_n))}{\prod_{i=1}^n f_i(F_i^{-1}(u_i))} = \frac{\phi_{\Sigma_W}(\Phi^{-1}(u_1),...,\Phi^{-1}(u_n))}{\prod_{i=1}^n \phi(\Phi^{-1}(u_i))}$

where $u_i = \Phi(W_i)$ and $F^{-1}(u) = \Phi^{-1}(u)$.

and $W_i = \Phi^{-1}\big(\frac{\rho_i - 1}{2}\big)$.

*Hyperpriors.*

$$\tau_\rho^2 \sim IG(a_1, b_1),\ \lambda_\rho \sim Uniform[0, 1],\ \sigma^2 \sim IG(a_2, b_2)$$

**SVAR model Full conditional distributions**

Full conditional distribution of $\boldsymbol{\beta}$ :

Given prior $\boldsymbol{\beta}|\tau_\beta^2 \overset{iid}{\sim} N(0, \tau_\beta^2)$, $[\boldsymbol{\beta}|\boldsymbol{Z}, \cdot] \propto [\boldsymbol{Z}|\mu, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}][\boldsymbol{\beta}|\tau_b^2]$

$\propto exp\big\{ -\frac{1}{2}[\boldsymbol{Z} - X\boldsymbol{\beta}]'\Sigma_Z^{-1}[\boldsymbol{Z} - X\boldsymbol{\beta}]\big\} \times exp\big\{ -\frac{1}{2\tau_\beta^2}\boldsymbol{\beta}'\boldsymbol{\beta}\big\}$

$\propto exp\big\{ -\frac{1}{2}[\boldsymbol{\beta}'X'\Sigma_Z^{-1}X\boldsymbol{\beta} + \boldsymbol{\beta}'X'\Sigma_Z^{-1}\boldsymbol{Z} + \boldsymbol{Z}'\Sigma_Z^{-1}X\boldsymbol{\beta} + \frac{1}{\tau_\beta^2}\boldsymbol{\beta}'\boldsymbol{\beta}]\big\}$

by completing the square, the full conditional is $\beta|\boldsymbol{Z},\cdot \sim MVN\Big(\mu_\beta, \Sigma_\beta\Big)$

$\Sigma_\beta = [X'\Sigma_Z^{-1}X + \frac{1}{\tau_\beta^2}]^{-1},\ \mu_\beta = \Sigma_\beta X'\Sigma_Z^{-1}\boldsymbol{Z}$

Full conditional of $\sigma^2$ :

We can write $\Sigma_Z = \sigma^2\Sigma_Z^*$

Given prior $\sigma^2|a_2, b_2 \sim IG(a_2, b_2),\quad [\sigma^2|\boldsymbol{Z},\cdot] \propto [\boldsymbol{Z}|\mu, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\rho}][\sigma^2|a_2, b_2]$

$\propto (\sigma^2)^{-nT/2}exp\big\{-\frac{1}{2\sigma^2}(\boldsymbol{Z}-\mu_Z)'{\Sigma_Z^*}^{-1}(\boldsymbol{Z}-\mu_Z)\big\} \times (\sigma^2)^{-a_2-1}exp\big\{\frac{-b_2}{\sigma^2}\big\}$

The full conditional distribution becomes:

$$\sigma^2|\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\rho} \sim IG\big(\frac{nT}{2} + a_2, \frac{1}{2}(\boldsymbol{Z}-\mu_Z)'{\Sigma_Z'}^{-1}(\boldsymbol{Z}-\mu_Z) + b_2\big)$$

Full conditional of $\tau_\rho^2$ :

We can write $\Sigma_W = \tau_\rho^2{\Sigma_W^*}^{-1} = \tau_\rho^2[(1-\lambda_\rho)I + \lambda_\rho R]^{-1}$

Given prior $\tau_\rho^2|a_1, b_1 \sim IG(a_1, b_1),\quad [\tau^2|\boldsymbol{W}, \lambda] \propto [\boldsymbol{W}|\tau_\rho^2, \lambda_\rho][\tau_\rho^2|a_1, b_1]$

$\propto (\tau_\rho^2)^{-n/2}exp\big\{-\frac{1}{2\tau_\rho^2}\boldsymbol{W}'{\Sigma_W^*}^{-1}\boldsymbol{W}\big\} \times (\tau_\rho^2)^{-a_1-1}exp\big\{\frac{-b_1}{\tau^2}\big\}$

The full conditional distribution becomes:

$$\tau_\rho^2|\boldsymbol{W},\lambda_\rho \sim IG\big(\frac{n}{2}+a_1,\frac{1}{2}\boldsymbol{W}'{\Sigma_W^*}^{-1}\boldsymbol{W}+b_1\big)$$

**SVAR model MH sampling algorithms**

Sampling $\lambda_\rho$ at iteration $k$ :

1. Draw $\lambda^*$ from proposal $q(\lambda_\rho^*|\lambda_\rho^{(k-1)}) \sim N(\lambda_\rho^{k-1},s_\rho^2)\mathbb{1}\{0\le\lambda_\rho^*\le 1\}$

   where $q(\lambda_\rho^*|\lambda_\rho^{(k-1)}) = \frac{\phi\left(\frac{\lambda_\rho^*-\lambda_\rho^{(k-1)}}{\tau_\lambda}\right)}{\Phi\left(\frac{1-\lambda_\rho^{(k-1)}}{s_\rho}\right)-\Phi\left(\frac{-\lambda_\rho^{(k-1)}}{s_\rho}\right)}$

2. Calculate acceptance ratio $r_\lambda = \frac{P(\lambda_\rho^*|\boldsymbol{Z},\boldsymbol{W}^{(k-1)},\cdot)q(\lambda_\rho^{(k-1)}|\lambda_\rho^*)}{P(\lambda_\rho^{(k-1)}|\boldsymbol{Z},\boldsymbol{W}^{(k-1)},\cdot)q(\lambda_\rho^*|\lambda_\rho^{(k-1)})} \propto \frac{[\boldsymbol{W}|\lambda_\rho^*][\lambda_\rho^*]\left[\Phi\left(\frac{1-\lambda_\rho^{(k-1)}}{s_\rho}\right)-\Phi\left(\frac{-\lambda_\rho^{(k-1)}}{s_\rho}\right)\right]}{[\boldsymbol{W}|\lambda_\rho^{(k-1)}][\lambda_\rho^{(k-1)}]\left[\Phi\left(\frac{1-\lambda_\rho^*}{s_\rho}\right)-\Phi\left(\frac{-\lambda_\rho^*}{s_\rho}\right)\right]}$

   $= \frac{|\Sigma_W(\lambda_\rho^*)|^{-1/2}exp\left\{-\frac{1}{2}\boldsymbol{W}'\Sigma_W^{-1}(\lambda_\rho^*)\boldsymbol{W}\right\}\left[\Phi\left(\frac{1-\lambda_\rho^{(k-1)}}{s_\rho}\right)-\Phi\left(\frac{-\lambda_\rho^{(k-1)}}{s_\rho}\right)\right]}{|\Sigma_W(\lambda_\rho^{(k-1)})|^{-1/2}exp\left\{-\frac{1}{2}\boldsymbol{W}'\Sigma_W^{-1}(\lambda_\rho^{(k-1)})\boldsymbol{W}\right\}\left[\Phi\left(\frac{1-\lambda_\rho^*}{s_\rho}\right)-\Phi\left(\frac{-\lambda_\rho^*}{s_\rho}\right)\right]}$

3. Generate $\xi \sim Uniform(0,1)$.

   $\lambda_\rho^{(k)} = \begin{cases}\lambda_\rho^* & if\ r_\lambda > \xi \text{ (with acceptance rate} = min\{1,r_\lambda\}),\\ \lambda_\rho^{(k-1)} & o.w.\end{cases}$

Sampling $\boldsymbol{\rho}$ at iteration $k$ :

1. Draw $\boldsymbol{\rho}^*$ from proposal distribution: $q(\boldsymbol{\rho}^*|\boldsymbol{\rho}^{(k-1)}) \sim C_{\Sigma_W^{**}}(\boldsymbol{\rho}^*|\boldsymbol{\rho}^{(k-1)})$ with $\Sigma_W^{**} = \tau_\rho^{*2}[(1-\lambda_\rho^*)I+\lambda_\rho^*R]^{-1}$ where $\tau_\rho^{*2}$ and $\lambda_\rho^*$ are proposal variance and spatial correlation parameters.

(a) Sample $\boldsymbol{W}^* \sim N(\boldsymbol{W}^{(k-1)}, \Sigma_W^{**})$.

(b) Transform $\boldsymbol{\rho}^* = 2\Phi(\boldsymbol{W}^*) - 1$.

2. Calculate acceptance ratio $r_\rho = \frac{P(\boldsymbol{\rho}^*|\boldsymbol{Z},\tau_\rho^{2(k-1)},\lambda_\rho^{(k-1)},\cdot)q(\boldsymbol{\rho}^{(k-1)}|\boldsymbol{\rho}^*)}{P(\boldsymbol{\rho}^{(k-1)}|\boldsymbol{Z},\tau_\rho^{2(k-1)},\lambda_\rho^{(k-1)},\cdot)q(\boldsymbol{\rho}^*|\boldsymbol{\rho}^{(k-1)})}$.

$$\propto \frac{[\boldsymbol{Z}|\boldsymbol{\rho}^*,\cdot][\boldsymbol{\rho}^*|\tau_\rho^{2(k-1)},\lambda_\rho^{(k-1)}]q(\boldsymbol{\rho}^{(k-1)}|\rho^*)}{[\boldsymbol{Z}|\boldsymbol{\rho}^{(k-1)}][\boldsymbol{\rho}^{(k-1)}|\tau_\rho^{2(k-1)},\lambda_\rho^{(k-1)}]q(\boldsymbol{\rho}^*|\boldsymbol{\rho}^{(k-1)})}$$

$$\propto \frac{|\Sigma_Z(\boldsymbol{\rho}^*)|^{-1/2}exp\{-\frac{1}{2}(\boldsymbol{Z}-\mu_Z)'\Sigma_Z^{-1}(\boldsymbol{\rho}^*)(\boldsymbol{Z}-\mu_Z)\}}{|\Sigma_Z(\boldsymbol{\rho}^{(k-1)})|^{-1/2}exp\{-\frac{1}{2}[\boldsymbol{Z}-\mu_Z]'\Sigma_Z^{-1}(\boldsymbol{\rho}^{(k-1)})[Z-\mu_Z]\}} \cdot \frac{f(\boldsymbol{\rho}^*)}{f(\boldsymbol{\rho}^{(k-1)})} \cdot \frac{q(\boldsymbol{\rho}^{(k-1)}|\boldsymbol{\rho}^*)}{q(\boldsymbol{\rho}^*|\boldsymbol{\rho}^{(k-1)})}.$$

Let $\Phi^{-1}(\frac{\boldsymbol{\rho}_i^*+1}{2}) = \boldsymbol{W}_i^*$ and $\Phi^{-1}(\frac{\boldsymbol{\rho}_i^{(k-1)}+1}{2}) = \boldsymbol{W}_i^{(k-1)}$.

Then, $f(\boldsymbol{\rho}^*) = \frac{\phi_{0,\Sigma_W}(W_1^*,\ldots,W_n^*)}{\prod_{i=1}^n \phi(W_i^*)} = \frac{exp(-\frac{1}{2}W^{*\prime}\Sigma_W^{-1}W^*)}{\prod_{i=1}^n \phi(W_i^*)}$

where $\Sigma_W = {\tau_\rho^{(k-1)}}^2[(1-\lambda_\rho^{(k-1)})I + \lambda_\rho^{(k-1)}R]^{-1}$.

and $q(\boldsymbol{\rho}^*|\boldsymbol{\rho}^{(k-1)}) = \frac{1}{2}g\big(\frac{\boldsymbol{\rho}^*+1}{2}\big)$

$$= \frac{\phi_{\Sigma_W}\left[(\Phi^{-1}(\frac{\rho_1^*+1}{2}),\ldots,\Phi^{-1}(\frac{\boldsymbol{\rho}_n^*+1}{2}))-\phi_{\Sigma_W}[(\Phi^{-1}(\frac{\rho_1^{(k-1)}+1}{2}),\ldots,\Phi^{-1}(\frac{\boldsymbol{\rho}_n^{(k-1)}+1}{2}))\right]}{2\prod_{i=1}^n \phi[\Phi^{-1}(\frac{\boldsymbol{\rho}_i+1}{2})]}.$$

3. Generate $\xi \sim Uniform(0,1)$.

$$\boldsymbol{\rho}^{(k)} = \begin{cases} \boldsymbol{\rho}^* & if\ r_\rho > \xi \text{ (with acceptance rate} = min\{1, r_\rho\}), \\ \boldsymbol{\rho}^{(k-1)} & o.w. \end{cases}$$